The Dissertation Committee for Charles Clement Traverse III certifies that this is the approved version of the following dissertation:

Genome-wide Detection of Transcription Errors in Bacteria

Committee:

Howard Ochman, Supervisor

Nancy Moran

Jeffrey Barrick

James Bull

Genome-wide Detection of Transcription Errors in Bacteria

by

Charles Clement Traverse III

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

December 2018

To my family and friends.

Acknowledgements

First and foremost, thank you to my advisor, Howard Ochman who gave me the opportunity to work in his lab, the support to complete this dissertation work, and the invaluable knowledge and skills he imparted to me. I would like to thank Dr. Louis-Marie Bobay for always critically reviewing my work, teaching me valuable skills, and for his continued willingness to discuss my ideas. I would like to thank to the entire Moran-Ochman lab for continued support and friendship. Thank you to Eli Powell and Kim Hammond for always making sure everything ran smoothly and efficiently. Finally, I would like to thank my parents and my girlfriend India for their undying love, support, and patience throughout my graduate studies.

Abstract

Genome-wide Detection of Transcription Errors in Bacteria

Charles Clement Traverse III, Ph.D.

The University of Texas at Austin, 2018

Supervisor: Howard Ochman

Errors in DNA that occur during replication serve as the basis for adaptation and heritable genetic variation in all organisms. However, non-heritable genetic variation will arise through errors that occur during transcription. Although it has been hypothesized that errors in transcripts might aid in survival of antibiotic stress and help evade immune responses, they can be deleterious in that they cause RNA polymerase (RNAP) to pause, resulting in collisions between the RNAP and replication proteins. Additionally, too many errors within the proteome can result in aggregation of these faulty proteins, inducing the general stress response. To avoid such complications, bacteria have evolved numerous mechanisms to improve the fidelity of transcription. These mechanisms include recognition of mis-paired bases within the RNAP, recognition of slippage along the template, excision of errors from the transcript, and prevention of errors from occurring.

Despite decades of research, accurate measurements of the transcription error rate have remained elusive. Although recent sequencing-based measurements can provide

ways to assay errors genome-wide, RNAseq error rates are too high to gauge the various types of substitutions. Additionally, measurement of transcription slippage remains limited to the analysis of long homopolymeric repeats in reporter genes. This dissertation reports the use of a sequencing technique that allows us to detect transcription errors and remove the errors that arise during sequencing. This method was successful in determining the genome-wide transcription substitution, insertion, and deletion rates. We found that transcription error rates remain constant across a wide range of growth states and across phylogenetically diverse bacteria. These data also suggest that transcription slippage occurs in sequences that are more complex than homopolymeric runs. Finally, we find that only one of the three previously identified transcription fidelity factors appears to influence transcription fidelity.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 OVERVIEW

Mutations in DNA are the source of heritable genetic variation in all organisms. However, non-heritable mutations in transcripts can also occur during the process of transcription. In humans, these errors have been associated with aging and the development of cancer (1). In bacteria, such errors occur approximately 10,000-fold more frequently than mutations to DNA (2–6) and are therefore pervasive across the transcriptome. Although these errors are transient in nature, they contribute significant variation to the proteome (7–9). Because these errors are likely to disrupt protein function, bacteria have evolved quality control mechanisms that serve to limit the occurrence of transcription errors (10). Historically, transcription errors have been measured using *in vitro* transcription assays or *in vivo* reporter genes (4–6, 11). Despite these efforts, the approaches fall short in measuring the transcription error rate in multiple ways because: *(i)* These measurements are indirect estimates of the transcription error rate, *(ii)* they are blind to genome-wide effects, *(iii)* they cannot differentiate between the different types of base substitutions, and *(iv)* they require separate assays to measure insertions and deletions (indels) (12).

This dissertation commences with a comprehensive background of the process of transcription and the biological facets of transcription errors in Chapter 1. Chapter 2 focuses on the application of a genome-wide sequencing technique that mitigates sequencing artefacts, termed CirSeq, to measure transcriptional errors (13, 14). Chapter 3

describes how indels can also be measured with CirSeq and the mechanistic insights that can be gained from indel patterns. Chapter 4 pivots to classifying the extent to which three transcription fidelity proteins prevent transcription errors. And chapter 5 brings the dissertation to an end with conclusions and future directions.

1.2 THE STRUCTURE OF THE RNA POLYMERASE AND TRANSLATION COUPLING

1.2.1 RNA polymerase

The RNA polymerase (RNAP) consists of five protein subunits that come together to form the 'core' enzyme (15). The core enzyme contains two identical α subunits, a β subunit, a β' subunit, and an ω subunit (15). This active core enzyme comes together with one of many possible σ factors to form the RNAP holoenzyme (15). The holoenzyme is capable of binding to stretches of DNA, called promoters, and initiating transcription (16).

A high resolution structure of the *E. coli* RNAP holoenzyme was notoriously difficult to solve. It remained recalcitrant until 2013 with the publication if the first high resolution *E. coli* RNAP holoenzyme at 3.6-Å resolution (17). This advancement allowed researchers to directly study the *E. coli* RNAP, as researchers had relied on structures of the RNAP from *Thermus* species until this point (15).

The two largest subunits of the RNAP core enzyme are the β and β' subunits (18). These two proteins bind together and guide the DNA template through the RNA

polymerase, maintain the separation of the two DNA strands, and catalyze the addition of new ribonucleotides to the growing transcript (19, 20). New ribonucleotides enter through a pore within these two proteins called the secondary channel (21). This channel also serves as a binding location for different regulatory proteins and molecules (10).

The two identical α subunits perform multiple roles for the RNAP. The N-terminal domains of α help assemble and stabilize the RNAP (22). Additionally, the flexible C-terminal domains of α aid in the initiation of transcription by interacting with certain promoters or upstream DNA motifs (23). After transcription has started, the C-terminal domains of these α subunits interact with the leading ribosome to aid in transcription-translation coupling (24).

The ω protein aids in the assembly of the RNAP core enzyme and helps maintain the stability of the RNAP throughout transcription (25, 26). The ω also has a regulatory role of responding to the guanosine pentaphosphate (ppGpp) molecule, which regulates the amino acid starvation response (stringent response) (25, 27).

Finally, σ factors bind to the RNAP core enzyme to form the RNAP holoenzyme (28). The σ factor is the part of the RNAP that recognizes promoters in DNA, which signify the RNAP to bind and initiate transcription (19). There are many types of σ factors that specifically bind to different promoter sequences, allowing for regulatory control over different genes in the genome (29). The σ factor is primarily involved in the initiation of transcription and unbinds from the RNAP core enzyme in a stochastic manner after transcription elongation has started (30).

1.2.2 Physical coupling of transcription and translation

In stark contrast to eukaryotes, where transcription and translation occur in separate cellular compartments, transcription and translation occur simultaneously in prokaryotes (31). It has been long known that transcription and translation co-occur in bacteria, and this is referred to as transcription-translation coupling. However, the extent of this coupling was not realized until 2010. In the same issue of *Science*, two separate publications from different lab groups reported that the ribosome controls the rate of transcription and that the ribosome and RNAP may be physically attached together (32, 33).

The mechanism of this physical interaction was thought to be mediated by the direct connection between the transcription elongation factor, NusG, and the ribosomal protein, NusE (34). This led to the view that the RNAP and ribosome were physically tethered together by these two proteins. A subsequent study demonstrated that the RNAP tended to pause and completely stop transcribing more frequently when this interaction was interrupted (35). This uncoupling also leads to increased genome instability by generating double-strand breaks, which are thought to be mediated by collisions between the paused RNAP and upstream, actively transcribing RNAPs (36, 37).

More recently, a new view of transcription-translation coupling arose that challenges the NusG:NusE bridging mechanism. A single particle cryo-electron microscopy (cryo-EM) study revealed that the physical connection between the ribosome and the RNAP was much more extensive than previously thought (24). In this new model, the nascent transcript leaves the RNAP and is immediately threaded directly into

the leading ribosome for translation. The authors termed the combined structure the 'expressome' and the structure shows many direct interactions between the ribosome and RNAP, indicating that they act as a single transcription/translation macromolecule. However, this newer model is in direct conflict with the NusG:NusE model because, according the cryo-EM structure, NusG and NusE are too far apart from one another to physically interact. Evidence for and against both mechanisms continues to arise (38–40), therefore the two models remain under active investigation. Another possibility is that both models are correct and they merely represent two different states of physical interaction that occur at different times (38, 40).

Regardless of which model, or if either, is correct, the interaction between the ribosome and RNAP is important for the process of transcription as a whole. However, this raises questions about the process of transcription in non-translated genes, those including tRNA or rRNA. If the ribosome mediates both the speed and pausing behavior of the RNAP in protein coding genes, is transcription fundamentally different in non-protein coding genes? Transcription does indeed operate differently in non-protein coding genes as discussed in the Section **1.3.4**.

1.3 TRANSCRIPTION

1.3.1 Initiation

Once the RNAP holoenzyme has assembled, it can interact with promoter sequences in the DNA template. Two domains in the σ subunit interact with the −10 and

−35 promoter elements, which are two conserved sequence motifs that are positioned 10 and 35 nucleotides upstream from the transcription start site, respectively (28). As stated above, the different σ proteins interact with different promoters, which correspond to nucleotide differences in the −10 and −35 promoter regions (29). After the RNAP has stably bound to the promoter, transcription initiation can begin.

Initiation starts by melting the DNA at the site that the RNAP is bound. This converts the 'closed' complex, where the RNAP is bound to double-stranded DNA, to the 'open' complex, where the DNA is unwound by the RNAP and the template strand enters the RNAP (41). The unwound DNA is also referred to as the 'transcription bubble'. The RNAP then starts transcription by adding the initiating nucleotide at the transcription start site (TSS or +1 site) (42). Unlike DNA polymerase, which requires a free 3'-OH group to start replicating DNA, the RNAP does not need a free 3'-OH on a primer and can initiate transcription by placing the ribonucleotide opposite to the template strand. At this point, the RNAP continues to add nucleotides to the 3' end of the nascent transcript.

The interactions of the σ subunit with the promoter sequence are strong, posing an energetic barrier to active elongation (42, 43). The RNAP will generate many successive short transcripts before aborting them and sliding back into its initial position at the TSS, a process referred to as 'abortive transcription' (44). Eventually, the RNAP will stochastically overcome the energetic barrier to allow for 'promoter escape' and make the transition from initiation into elongation (42–45).

1.3.2 Elongation

Once the energetic barrier to promoter escape has been overcome, new ribonucleotides will be added to the transcript in a processive manner. During transcription elongation, the most recently transcribed base resides in the *i* position and the template DNA to be transcribed next resides in the *i*+1 position (46). The incoming ribonucleotide then base pairs with the template in the *i*+1 position and is stabilized by conserved amino acid residues (46). The addition of the new ribonucleotide to the growing transcript is mediated by a catalytic Mg^{2+} ion and the then translocation of the RNAP takes place, wherein the whole RNAP moves forward by one base along the template DNA (47). Translocation moves the newly transcribed ribonucleotide from the *i*+1 to the *i* position, allowing for the next cycle of ribonucleotide addition to take place.

During elongation, nine of the most recently transcribed nucleotides remain hybridized to the DNA template (48). As each new base is added to the 3' end of the transcript, the 5' end of this RNA:DNA hybrid dissociates from the template DNA and enters the transcript exit channel (49). Proper base pairing in this RNA:DNA hybrid is referred to as the register, and maintaining this register is important for both elongation processivity and RNAP stability (48, 50, 51).

1.3.3 Termination

There are two mechanisms for terminating transcription in *E. coli*: intrinsic termination and Rho termination (52). Intrinsic termination occurs without the help of any additional proteins, but is instead mediated by the transcript itself. Intrinsic

terminators contain a stem loop, which is formed by a G+C-rich inverted repeat that hybridizes together to form a stable secondary structure (53). Following the inverted repeat, a string of rU nucleotides are transcribed, ensuring that the RNA:DNA hybrid will be occupied by weak rU:dA pairs (53). The formation of the stem loop imposes steric constraints on the RNA:DNA hybrid, causing the 5' bases in the RNA:DNA hybrid to dissociate from the template (52–54). Because this RNA:DNA hybrid is weakened from the of rU:dA pairs, complete dissociation of the remaining hybrid readily occurs (52, 53).

The second mechanism for transcription termination involves a protein that translocates along the transcript and physically interrupts transcription. This Rho protein is responsible for 20–30% of all termination events in *E. coli* (55, 56). Unlike intrinsic terminators, which are easy to identify bioinformatically, Rho-dependent terminators are hard to identify because there are no specific sequence motifs that signify Rho will bind (57). Despite the lack of specific sequence motifs, pyrimidine-rich regions, usually C, within a 60–90 base region are the canonical binding sites for Rho and are termed Rho utilization sites (*rut*) (52, 57). Once the *rut* site is bound to all six Rho subunits, the RNA is looped through the center of Rho and ATP is used to power the translocation of Rho along the RNA and toward the elongating RNAP (52, 57). Once Rho collides with RNAP, the transcript is pulled out from the RNAP and transcription is effectively terminated (52, 57).

Recent evidence has shown that Rho associates with actively elongating RNAP and waits for an opportunity to bind to the transcript (30). The exact mechanism for this process is still under investigation, but transcription-translation coupling is known to

block access of the transcript from Rho (32, 52). And under the recent expressome model, this blocking may occur because the transcript immediately enters the ribosome right after it leaves the RNAP exit channel (24). Only after translation is terminated or uncoupled from transcription, will Rho have an opportunity to bind to the transcript for termination (52).

1.3.4 Antitermination

If rRNA and tRNA genes are not translated, then how do they overcome the increased Rho binding and RNAP pausing behavior that should emerge without a leading ribosome to inhibit these activities? To solve this problem, a series of proteins interact with RNAPs that initiate transcription at rRNA and tRNA promoters (58, 59). This protein complex is made up of NusA, NusB, NusE, NusG, among other proteins, and associates at the RNAP exit channel (58). This effectively allows the RNAP to transcribe through *rut* sites without Rho terminating transcription (58, 59). The mechanism behind antitermination is, in part, a marked increase in the transcription elongation rate from 45 nucleotides per second to 65 nucleotides per second (60, 61). This increase in the rate of elongation allows for secondary structure to form quickly enough sequester *rut* sites and block Rho from binding (58, 59). Additionally, a faster elongation rate allows the RNAP to bypass intrinsic terminators (58).

1.4 TRANSCRIPTION MISINCORPORATIONS AND ERROR CORRECTION

As important as transcription is to the cell, RNAP does make mistakes. When the wrong nucleotide is transcribed, generating a mismatch, the hydrogen bonds between the mismatched nucleotide and the template DNA do not properly align. This induces an incorrect conformation in the RNA, and the mismatch will "fray" off of the template and induce the RNAP to pause transcription (50, 51). The RNAP will then translocate backwards along the DNA template and extrude the mismatched nucleotide out of the active *i* site and into a proofreading site, a process referred to as backtracking (46, 62). After backtracking by 1–2 bases, the mismatch can be cleaved from the transcript and transcription can resume (46, 63). There are two known mechanisms for inducing the cleavage of mismatched bases: intrinsic cleavage and Gre-mediated cleavage (10).

1.4.1 Intrinsic cleavage

This first mode of transcript cleavage is internal and intrinsic to the RNAP itself. The same active site and Mg^{2+} ions that catalyze the addition of each new nucleotide to the growing transcript can also induce the cleavage of backtracked nucleotides (46, 63). After the mismatched nucleotide has been backtracked into the proofreading site, the transcript is cleaved and the mismatched base, sometimes along with the base immediately before it, is ejected through the secondary channel (46, 63). *In vitro* experiments that measure intrinsic cleavage show it to be a slow and inefficient process, thus leaving its biological contribution to error correction *in vivo* unknown (46, 63).

1.4.2 Gre-mediated cleavage

The second mechanism of transcript cleavage involves the GreA and GreB proteins. Both of these proteins bind to the secondary channel of the RNAP and can stimulate transcript cleavage by stabilizing the Mg^{2+} ion that performs the cleavage activity (64). Gre-medated cleavage is faster than intrinsic cleavage and is thought to be the major mechanism for correcting transcription misincorporations (10, 63). Whereas these proteins correct errors during transcription, they also serve as general anti-backtracking factors (36).

The RNAP can backtrack stochastically or in response to a misincorporation, and these proteins ensure that backtracking will be resolved quickly to avoid any RNAP-RNAP or DNAP-RNAP collisions (36). GreA and GreB, although structurally and functionally similar, have differing roles in the cell (10). GreA has been shown to associate with short 1–3 base backtracking events, whereas GreB stimulates transcript cleavage of longer backtracking events of up to 18 bases in length (65–69).

1.4.3 DksA-mediated error prevention

A third protein, DksA, which is structurally similar to the Gre proteins and binds to the same site on the RNAP, has recently been classified as a transcription fidelity factor (10). However, DksA has long been known to regulate and re-program the RNAP to induce the stringent response, making transcription fidelity a new role attributed to this protein (70, 71). In conjunction with the ppGpp alarmone, DksA inhibits the synthesis of ribosomes during amino acid starvation but also stimulates transcription initiation at

amino acid synthesis promoters (70–72). The mechanism for this seemingly contradictory pattern is thought to operate by DksA/ppGpp lowering the energetic barrier to open complex formation while simultaneously reducing the stability of the open complex (73, 74). The sequence differences between the two promoter types are thought to result in differing outcomes: transcription initiation of rRNA is inhibited, whereas transcription initiation of amino acid synthesis genes is stimulated (72). The precise reason for the differing behavior of the two promoter types is still under active investigation.

DksA was historically thought to only act during transcription initiation, but more recent evidence shows that DksA is bound to RNAP throughout transcription elongation, hinting at a role during this phase of transcription (35). Following this finding, DksA was shown to inhibit misincorporation events *in vitro* (75). Additionally, an *in vivo* study used two different reporter assay systems to demonstrate a role for DksA in reducing misincorporation events (76). The mechanism for this reduction in transcription errors remains a mystery because DksA does not induce nucleotlytic cleavage, potentially representing a new pathway for reducing transcription errors

1.5 TRANSIENT ERRORS: OF WHAT CONSEQUENCE ARE NON-HERITABLE MUTATIONS IN TRANSCRIPTS AND PROTEINS?

1.5.1 Subclasses of transient errors

Despite the quality control inherent to transcription and translation, there are errors that escape these mechanisms. Transient errors originate through mistakes in

information transfer along the pathway from DNA to protein. They are considered transient because both transcripts and proteins are short-lived; therefore any errors in the two will not be heritable. Errors can occur at each step in this pathway, including during transcription, translation, and loading of the incorrect amino acids onto tRNAs (77). Each of these errors can have different effects for the cell and they each occur at different rates. Transient errors have been historically difficult to measure, usually relying on indirect measurements using reporter genes.

The first point where transient errors can be introduced into the process of gene expression is during transcription. According to reporter gene assays, these errors are thought to occur at a rate of 10^{-4} to 10^{-5} per nucleotide (4, 5). However, not much is known about error rates across the genome, the rates of individual substitutions, or the indel rates (2, 12, 77). Because transcription errors are introduced early in gene expression, any error occurring in a transcript will serve as the template for translation and will be translated into protein. Additionally, a single transcript can be translated up to 40 times in bacteria, ensuring that each singular transcription error will be present in each protein that arises from that transcript, assuming the error results in an amino acid change (78, 79).

The other type of transient error occurs during translation. There are two main sources of these errors: from the ribosome itself and from tRNAs containing the incorrect amino acid (80, 81). The combined effects of these two sources of error occur at a rate of ~10^{-3}, with different rates for each amino acid (82). Although translation errors are 1–2

orders of magnitude more frequent than transcription errors, the individual effect of each translation error is smaller because only one amino acid is changed in one protein.

The fact that one deleterious transcription error can be amplified into a burst of flawed protein may explain why the transcription error rate is lower than the translation error rate: a single deleterious transcription error could have a much larger effect than a single deleterious translation error. Additionally, those transcription errors that are recognized by the RNAP will induce stalling and backtracking, which can lead to double strand breaks if they are not resolved quickly (36). Taken together, each individual transcription error can induce more negative consequences to the cell than each translation error, making them more important to control.

1.5.2 Phenotypic consequences of transient errors

The physiological effects of transient errors are poorly understood and two perspectives are commonly explored in the literature: that transient errors can be beneficial or that transient errors are more harmful than beneficial (7, 83). The argument that these transient errors can be beneficial stems from the idea that increasing the diversity of the proteome within the population could allow a sub-population of cells to survive a sudden stress (7, 8, 84). If the population is not genetically equipped to handle a particular stress or environment, then consistently generating extensive, but temporary, diversity could end up being more beneficial than restricting these errors from occurring in the first place.

The benefits of this temporary diversity could manifest themselves in multiple ways. For example, an amino acid substitution in a metabolic enzyme could allow that enzyme to process a molecule that would not normally bind to that enzyme (7). Alternatively, transient errors in a regulatory enzyme could result in the sudden activation of an operon that would not normally be induced (85). And, by extension, these errors could change the regulatory landscape if a genetic program such as the general stress response or stringent response were induced that pre-equipped these cells to handle a sudden stress without needing to wait for their induction through traditional mechanisms (84). Finally, there is evidence that increasing the rate of transient errors results in the accumulation of many faulty proteins. Although this may be detrimental to the cell, the accumulation of these proteins induces the general stress response and increases the resistance of the population to challenges like oxidative stress (84).

However, the extent that these errors are beneficial to the population remains limited to specialized, laboratory-derived examples and speculation. Transcription errors are known to result in double strand breaks when DNA replication enzymes encounter stalled RNAP if they are not corrected in time (36). Additionally, an increase in transcription errors in yeast has been associated with a reduced cellular lifespan (83). Therefore, it is more likely that transient errors, particularly transcription errors, are more detrimental to the cell than beneficial.

1.6 HOW HAVE TRANSCRIPTION ERRORS BEEN MEASURED?

1.6.1 Radiolabeled *in vitro* transcription assays

The initial estimates of bacterial transcription fidelity were based on the rate that radiolabeled ribonucleotides were erroneously misincorporated into RNA that was transcribed from repeating dinucleotide templates (11). Springgate and Loeb (1975) measured misincorporation of radiolabeled rC or rG into poly-dAdT chains and rA or rU into poly-dCdG chains using purified *E. coli* RNAP core and holoenzymes. They were able to determine that each substitution type had different rates, but the transcription assay approach suffers from multiple problems. For example, repeating templates are known to induce errors in both DNA and RNA polymerases (12, 86). Additionally, the core and holoenzymes were tested without additional elongation factors, transcription-translation coupling, transcription fidelity factors, or regulatory molecules. However, this study marks the first measurement of transcription errors and represents a milestone in transcription fidelity research.

1.6.2 LacZ nonsense alleles

The first *in vivo* transcription error measurements were performed using a *lacZ* reporter gene with a premature stop codon engineered early into the gene (4, 5). The assay worked by allowing the cells to grow after LacI induction and then measuring the β-galactosidase activity of LacZ. Only cells that experienced a transcription error that converted the premature stop codon to an amino acid would be able generate a functional

LacZ tetramer. They used a highly polar *lacZ* mutation that they argued would mitigate the effects of translation errors in this system because Rho would terminate the RNAP after the leading ribosome encountered the stop codon. In effect, this meant that no full-length transcripts would be generated that could then be mis-translated into functional LacZ. The authors argued that the translation error rate was so low that translation errors would be negligible to the system. Furthermore, a single transcription error is sufficient to generate multiple functional LacZ tetramers because each transcript that is corrected by a transcription error will be translated multiple times. In contrast, it would take four successive translation errors in a single cell to complete a functional LacZ tetramer.

However, the error rates derived from this study relied on extensive back-calculation from the β-galactosidase signal that required many assumptions. This back-calculation (*i*) required precise measurements of the number of input cells and β-galactosidase activity, (*ii*) relied on estimates for the number of functional LacZ tetramers that arose from each individual transcription error, and (*iii*) needed to know how many functional LacZ molecules were successfully extracted from the input cell pool. Together, the authors calculated a transcription error rate of ~1.4 x 10^{-4}. Whereas this measurement served as the first measurement of the *in vivo* transcription error rate, more recent measurements of the translation error rate indicate that translation errors may occur frequently enough to influence the assay (8, 82). Although this assay may be influenced by translation errors, LacZ reporters are still used to measure the relative differences of transcription error rates in mutants.

1.6.3 Sequencing data

More recently, there have been efforts to measure the transcription error rate using high-throughput sequencing. High throughput sequencing on the Illumina platform is a technique that allows for millions of short sequences of DNA between 50–300 bases to be simultaneously sequenced. RNA can also be sequenced after a methodological step that converts the RNA into a DNA copy (cDNA). RNA sequencing (RNAseq) provides a promising method for measuring the transcription error rate across the entire genome. However, the error rate of RNA sequencing is higher than previous reports for the transcription error rate, making it impossible to determine if the source of an error arose during transcription or as an artefact of sequencing (87, 88).

Despite the high error rate of RNAseq, there are clever strategies that attempt to circumvent this problem. A technique developed in 2013 reported the transcription error rate in the *Caenorhabditis elegans* to be ~4 x 10^{-6} (89). The authors used an altered library preparation technique that ligates an eight nucleotide randomized barcode onto the 5' end of each RNA fragment. These eight nucleotides also contain a biotin at the 5' end, allowing for capture of each RNA fragment with streptavidin beads and a magnet. Following capture, three individual cDNA libraries are made from the same captured RNA source and the libraries are prepared for sequencing. Because each RNA fragment has eight random nucleotides associated with it, two or three copies of the same RNA fragment may be present in the sequencing data that can be matched together with the barcodes. This allows for a consensus sequence to be made among sequencing reads that are matched together. Although this method was successful in measuring the transcription

error rate, the efficiency is very low because it relies on the chance event that the same RNA fragment with the same barcode will be present in two or three of the library replicates.

Another group recently attempted to measure the transcription error rate using sequencing data from the Nascent Elongating Transcript Sequencing (NET-seq) protocol (90). The NET-seq protocol isolates RNAPs and then degrades all RNA that is not captured within the RNAP. After purification of the RNAPs, the only RNA present should be the RNA that was residing inside of the RNAP, representing the RNA in the RNA:DNA hybrid at the time the cells were harvested. The authors repurposed this dataset to measure the error rates in each position within the RNA:DNA hybrid of the RNAP. By applying rigorous quality control to the sequencing data, they were able to see differences in the error rates between the most recently transcribed nucleotide (3' most position in the RNA:DNA hybrid) and the rest of the nucleotides in the RNA:DNA hybrid. However, because standard RNAseq is so error-prone, this technique is more powerful for measuring relative differences in mutants rather than calculating absolute error rates because sequencing errors are prevalent in the data.

1.7 CIRSEQ: A METHOD TO MITIGATE SEQUENCING ARTEFACTS

Whereas this dissertation used two methods to measure the transcription error rate (see Chapter 2), only the method described in this section was capable of measuring transcription errors in our hands: CirSeq (13, 14). The description in this section is a conceptual overview of the technique, and the details of the method can be found later in

Chapter 2 in the Materials and Methods section. This RNAseq method was originally designed to ascertain the mutation rate of RNA viruses and adds modifications to RNAseq library preparation.

Traditional bacterial RNAseq protocols for the Illumina platform start with purified mRNA that is fragmented to smaller chunks of RNA. Random hexamers (six random nucleotides of DNA) are then added to the fragmented RNA and cDNA is synthesized using reverse transcriptase. Next, a second strand is synthesized from the cDNA, resulting in a double stranded DNA copy of the original mRNA fragment. Specific Illumina-specific adapters are then ligated onto the cDNAs, enabling them to bind to the Illumina sequencing chip. The libraries are then PCR amplified prior to sequencing.

The CirSeq protocol uses this same method, but introduces a few modifications and new key steps. First, the RNA is fragmented to an average of 80–100 nucleotides, run on a polyacrylamide gel, and RNA of the same size range is extracted from the gel. These fragments are then circularized using an RNA ligase and then primed with random hexamers. During the cDNA synthesis step, the reverse transcriptase will travel around the circularized RNA fragments many times, displacing any cDNA that it encounters. The resulting cDNA can reach lengths of tens of kilobases, so the cDNA is fragmented using sonic waves and then size selected for 300 nucleotides. After this size selection step, each cDNA will contain multiple linked repeats of the original mRNA fragment. To finish the libraries, Illumina adapters are ligated onto the cDNAs and then PCR

amplified. The libraries are then ready for sequencing on an Illumina MiSeq machine with a 1x300 read length kit.

Because 80–100 nucleotide fragments were used as the input RNA, most of the 300 base cDNAs should contain at least 3 repeats of the original mRNA fragments. After sequencing, a data pipeline developed by Acevedo *et al*. (2013) is used to process each read from linked repeats into a singular consensus sequence. Only reads containing three repeats are used for further analysis and the consensus sequence is generated using the quality scores of each base in each repeat. The method is summarized in Figure 1.1.

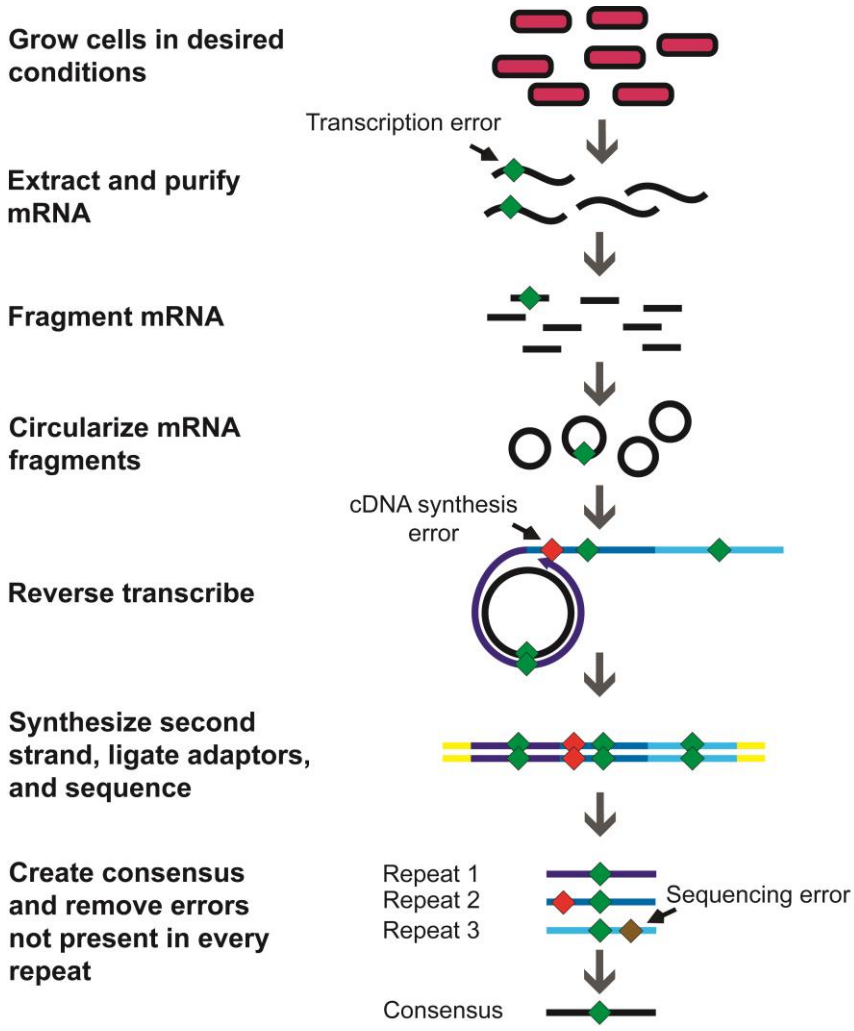

Figure 1.1 – Workflow of RNA circle sequencing

First, cells of interest are grown in the desired conditions and the RNA is extracted and purified. This mRNA is then fragmented, and sizes between 80 and 100bp are manually extracted from a polyacrylamide gel. Next, the fragments are circularized and subsequently reverse transcribed, resulting in linked cDNA repeats from the mRNA fragments. After second strand synthesis, the cDNA repeats are fragmented and sizes of 300bp are extracted. An Illumina library is then generated and sequenced on a MiSeq 1x300 run. The resulting sequences are then analyzed by the CirSeq pipeline (13). Only bases that differ from the reference genome within all three repeats are considered to be an error.

Chapter 2: Conserved rates and patterns of transcription errors across bacterial growth states and lifestyles[1]

2.1 ABSTRACT

Errors that occur during transcription have received much less attention than the mutations that occur in DNA because transcription errors are not heritable and usually result in a very limited number of altered proteins. However, transcription error rates are typically several orders of magnitude higher than the mutation rate. Also, individual transcripts can be translated multiple times, so a single error can have substantial effects on the pool of proteins. Implementing a method that captures transcription errors genome-wide, we measured the rates and spectra of transcription errors in *E. coli* and in endosymbionts for which mutation and/or substitution rates are greatly elevated over those of *E. coli*. Under all tested conditions, across all species, and even for different categories of RNA sequences (mRNA and rRNAs), there were no significant differences in rates of transcription errors, which ranged from 4.67×10^{-5} per nucleotide in mRNA of the endosymbiont *Buchnera aphidicola* to 5.09×10^{-5} per nucleotide in rRNA of the endosymbiont *Carsonella ruddii*. The similarity of transcription error rates in these bacterial endosymbionts to that in *E. coli* (8.23×10^{-5} per nucleotide) is all the more surprising given that genomic erosion has resulted in the loss of transcription fidelity factors in both *Buchnera* and *Carsonella.*

[1] This chapter is reproduced (with minor modifications) from its initial publication:

Traverse CC and Ochman H (2016) Conserved rates and patterns of transcription errors across bacterial growth states and lifestyles. *Proc Natl Acad Sci USA* 113:3311–3316. Ochman H supervised the project.

2.2 INTRODUCTION

Among the multiple types of information processing errors, the majority of research has focused on mutations that occur during DNA replication since such errors are heritable and form the basis of evolutionary change. However, errors that occur during transcription and translation can also have substantial effects on gene function by producing misfolded and malfunctioning proteins. The translation error rate is typically an order of magnitude higher than the rate of transcription errors (4, 5, 11, 91–93). However, errors occurring during transcription often elicit more dire consequences than those occurring during translation because individual mRNAs can be translated up to 40 times (78, 79), resulting in a burst of flawed proteins. Therefore, a single transcription error can result in many flawed proteins whereas a translation error will disrupt only a single protein.

Because deleterious transcription errors are not transmitted to subsequent generations, they can occur more frequently than mutations to DNA but still infrequent enough to ensure the cell is not overburdened with faulty proteins. Estimates of the rate of transcription errors in *E. coli* have been determined *in vitro* by measuring the misincorporation of radiolabeled ribonucleotides into repeating dinucleotide tracts (6, 11), and *in vivo* by quantifying the reversion frequencies of nonsense mutations in *lacZ* (4, 5). These assays yielded variable estimates of transcription error rates of 10^{-4} to 10^{-5} per nucleotide, several orders of magnitude higher than the mutation rate (94–96). Studies that assay individual loci are often not representative of the genome as a whole because sequence- or genome-specific features, such as base composition (96, 97) or sequence

motifs (12), affect the incidence of information processing errors. Moreover, transcription error reversion assays based on the recovery of functional proteins might also include translation errors, if these occur at a sufficiently high rate.

RNAseq offers an approach to both disentangle transcription errors from translation errors and provide an error rate for every transcribed gene in a genome. Unfortunately, the high error rates both of cDNA synthesis (3–6 x 10^{-5} per nucleotide; 100–102) and of high-throughput sequencing technologies (possibly as high as 10^{-2}–10^{-3} per nucleotide) (87, 88) renders the transcription errors obtained by conventional RNAseq indistinguishable from sequencing artefacts. Two recently developed methods offer ways to circumvent these problems by allowing transcription errors to be distinguished from sequencing and cDNA synthesis errors. Through the use of altered library preparation protocols, these methods reduce the predicted error rate of RNAseq to less than 10^{-8} (89) and 10^{-12} (13, 14) per nucleotide, making it possible to measure error rates across the entire transcriptomes of viruses and other organisms.

In this chapter, we describe and implement both of these RNAseq-based methods in *Escherichia coli* to examine if transcription error rates vary according to growth state and physiological condition, as has been reported for translation error rates (93, 101–103) and for the combined transcription and translation error rate (8). However, we found that only one of these methods were effective in measuring the transcription error rate (discussed later in this chapter). Moreover, we ask if transcription error rates are increased in the endosymbiotic bacteria *Buchnera aphidicola* and *Carsonella ruddii*—species that have lost known transcription fidelity factors and whose mutation rates,

substitution rates, and rates of protein sequence evolution are all amplified as a result of genetic drift and the loss of repair enzymes (Figure 2.1). We show that transcription error rates are remarkably similar across organisms, even for broad categories of RNA on which the cell is known to selectively degrade malfunctioning rRNA (104).

Figure 2.1

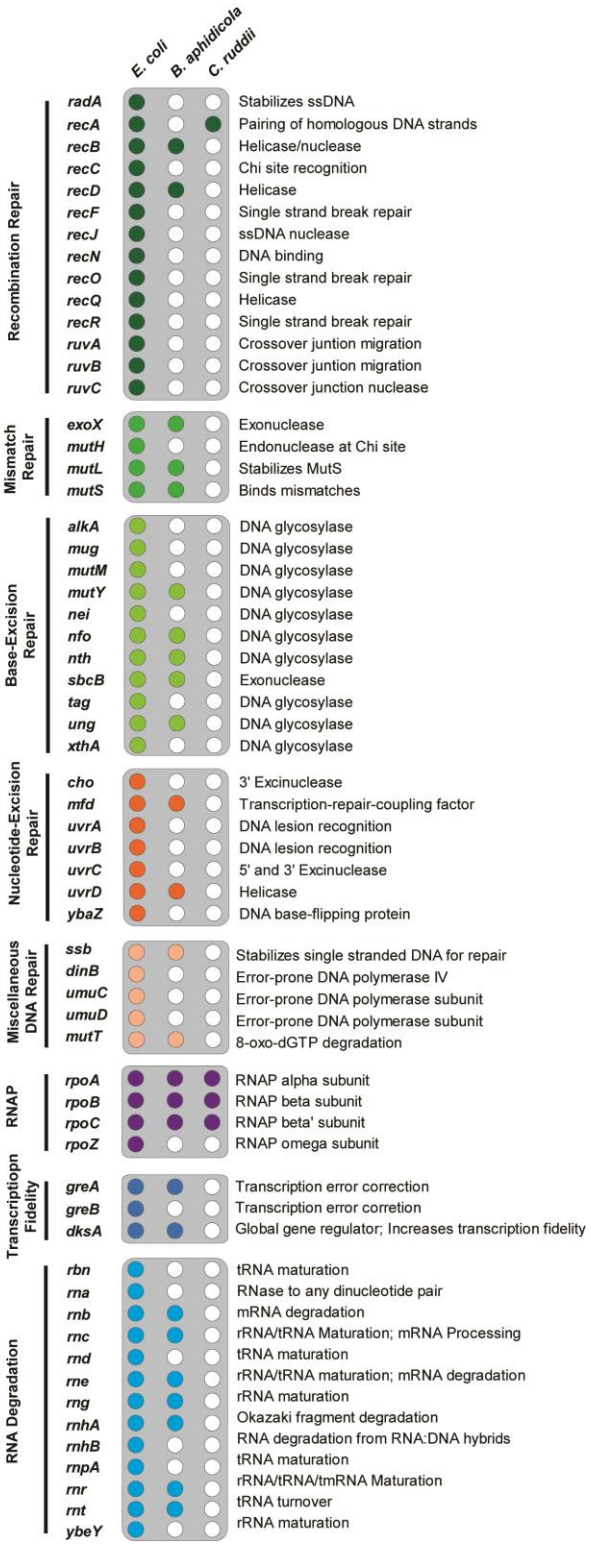
E. coli
B. aphidicola
C. ruddii
Recombination Repair
radA Stabilizes ssDNA
recA Pairing of homologous DNA strands
recB Helicase/nuclease
recC Chi site recognition
recD Helicase
recF Single strand break repair
recJ ssDNA nuclease
recN DNA binding
recO Single strand break repair
recQ Helicase
recR Single strand break repair
ruvA Crossover juntion migration
ruvB Crossover juntion migration
ruvC Crossover junction nuclease
Mismatch Repair
exoX Exonuclease
mutH Endonuclease at Chi site
mutL Stabilizes MutS
mutS Binds mismatches
Base-Excision Repair
alkA DNA glycosylase
mug DNA glycosylase
mutM DNA glycosylase
mutY DNA glycosylase
nei DNA glycosylase
nfo DNA glycosylase
nth DNA glycosylase
sbcB Exonuclease
tag DNA glycosylase
ung DNA glycosylase
xthA DNA glycosylase
Nucleotide-Excision Repair
cho 3' Excinuclease
mfd Transcription-repair-coupling factor
uvrA DNA lesion recognition
uvrB DNA lesion recognition
uvrC 5' and 3' Excinuclease
uvrD Helicase
ybaZ DNA base-flipping protein
Miscellaneous DNA Repair
ssb Stabilizes single stranded DNA for repair
dinB Error-prone DNA polymerase IV
umuC Error-prone DNA polymerase subunit
umuD Error-prone DNA polymerase subunit
mutT 8-oxo-dGTP degradation
RNAP
rpoA RNAP alpha subunit
rpoB RNAP beta subunit
rpoC RNAP beta' subunit
rpoZ RNAP omega subunit
Transcriptiopn Fidelity
greA Transcription error correction
greB Transcription error corretion
dksA Global gene regulator; Increases transcription fidelity
RNA Degradation
rbn tRNA maturation
rna RNase to any dinucleotide pair
rnb mRNA degradation
rnc rRNA/tRNA Maturation; mRNA Processing
rnd tRNA maturation
rne rRNA/tRNA maturation; mRNA degradation
rng rRNA maturation
rnhA Okazaki fragment degradation
rnhB RNA degradation from RNA:DNA hybrids
rnpA tRNA maturation
rnr rRNA/tRNA/tmRNA Maturation
rnt tRNA turnover
ybeY rRNA maturation

Figure 2.1 – Nucleic acid processing genes

Nucleic acid information processing genes that are present in *E. coli* compared to their retention or loss in *B. aphidicola* and *C. ruddii*. Colored circles indicate retention of the corresponding gene, white circles indicate loss of the corresponding gene from the specified genome.

2.3 RESULTS

2.3.1 Resource limitation and growth phase do not alter transcription error rates

We tested the effects of different growth conditions—all of which have been associated with altered mutation rates and/or translation error rates—on rates of transcription errors. Using a deep-sequencing approach to identify errors, we measured the transcription error rate in *E. coli* when grown under four growth conditions (TSB complex media or M9 minimal media, each sampled at mid-log and at stationary phase). Note that these errors include both base substitutions during the process of transcription and any damage to the mRNA after transcription. Each of the four conditions were assayed in duplicate, and in total, we detected 4,429 transcription errors, with the number of errors per sample ranging from 227 to 1,096. In neither of the nutrient sources were there significant differences in transcription error rates for cells harvested at mid-log phase or at eight hours after entering stationary phase (Figure 2.2, paired Wilcoxon test, $p > .30$). Similar to what we observed for *E. coli* assayed at different growth phases, transcription error rates do not differ significantly in nutrient-rich (TSB) and nutrient-poor (M9) growth media (Figure 2.2, paired Wilcoxon test, $p = .3429$). Furthermore, there are no significant differences in overall transcription error rates between any pair of individual conditions tested (Figure 2.2; two-tailed *t*-tests, $t(2) < 2.3$, $p > .14$); and the average transcription error rate over all conditions is 8.23×10^{-5} for *E coli* mRNA.

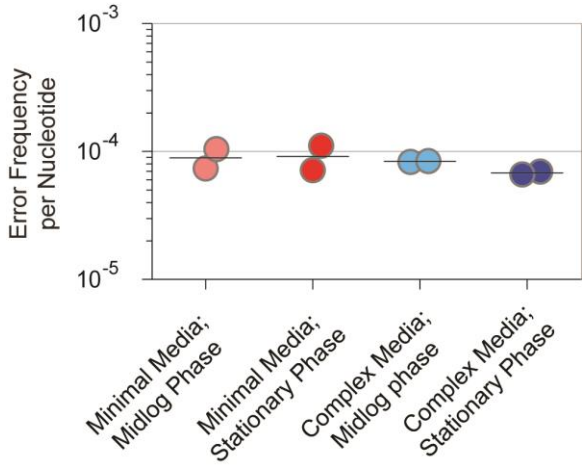

Figure 2.2 – Frequency of transcription errors in *E. coli*

Points are color-coded according to growth condition (n = 2 for each condition); horizontal bars represent means of each column. No significant differences in transcription error frequencies were detected between any of the tested parameters.

2.3.2 Distribution of transcription errors

The use of a high-throughput sequencing method to detect transcription errors (as opposed to a reporter-gene method) enables analysis of transcription errors genome-wide as well as the localization of errors to individual sites in each transcript. Starting at the scale of whole genomes, we analyzed the fluctuation in transcription error rates and found that the 95% of measurements made for 50-kb non-overlapping windows across the entire *E. coli* genome varies 3.5-fold among genomic regions, ranging from 3.3–10.1 x 10^{-5} (Figure 2.3). Regions containing highly expressed genes had an increased number of transcription errors (Figure 2.4), resulting from increased coverage enabling the discovery of more errors relative to areas in the genome with low coverage.

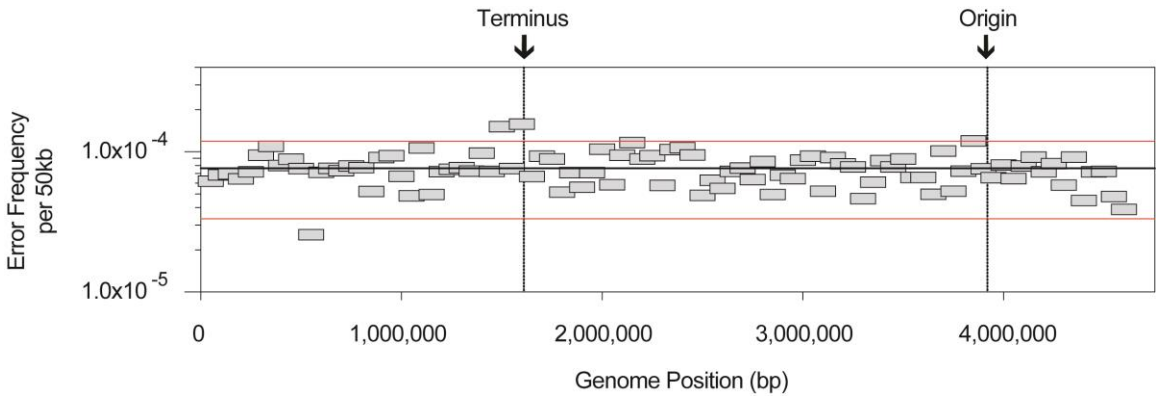

Figure 2.3 – Frequency of transcription errors along the *E. coli* genome

Shaded rectangles represent transcription error rates of all errors over the eight *E. coli* samples in non-overlapping 50-kb windows. Horizontal lines represent the genome-wide mean transcription error rate (black) and two standard deviations from the mean (red). Positions of replication origin and terminus are shown.

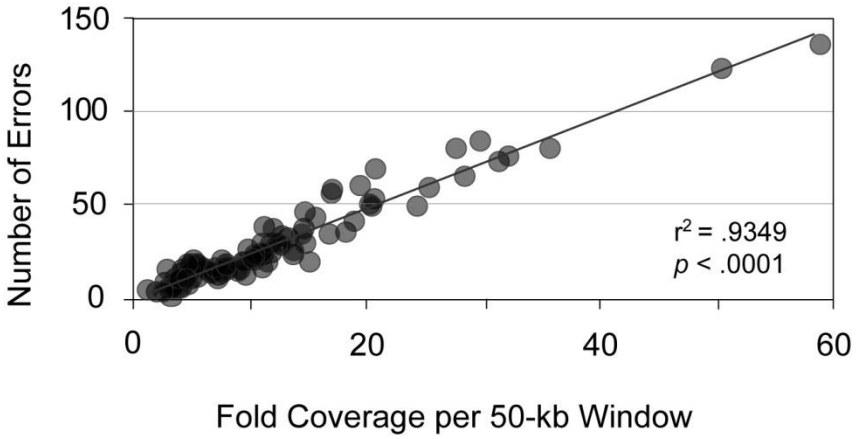

Figure 2.4 – Association between numbers of transcription errors and sequence coverage

Error numbers computed for non-overlapping 50-kb windows across the *E. coli* genome in all eight samples.

Transcription proceeds in the direction of DNA replication on the leading strand and in the opposite direction on the lagging strand, in which case there can be collisions between the replication and transcription machineries. Despite an increased likelihood of collision-induced errors on the lagging strand, there is no significant difference in the transcription error rates between genes encoded on the two strands (Wilcoxon test, $p > .90$). Next, we tested if adjacent nucleotides affected the occurrence of transcription errors and found that neither a particular preceding nor succeeding nucleotide induced transcription errors. Only when both the preceding and succeeding nucleotides are guanine residues do we observe a significant increase in transcription error frequency (Fisher's exact test, $p < .02$). Taken together, transcription errors occur without regard for genome location, direction of transcription, or for the vast majority of neighboring nucleotides.

2.3.3 Biases in *E. coli* transcription errors

Measuring transcription errors using a sequencing-based approach provides information about the absolute frequencies of each of the possible base substitutions. C→U errors were most common, occurring at a significantly higher frequency than all other transcription errors (Figure 2.5A), presumably attributable to high rates of cytosine deamination after the RNA is transcribed. It has previously been reported that transcription errors incur a higher rate of transitions than transversions (2, 89), the same overall pattern that we observe in *E. coli* (Wilcoxon test, $p < .05$). This trend, however, is driven solely by high incidence of C→U changes and no longer reaches significance after

removing these transitions from the analysis (Wilcoxon test, $p > .50$). Next, we tested the effect of individual nucleotides on the frequency of transcription errors in *E. coli* and found that G/C→N errors occur at higher frequencies than do A/U→N errors (Wilcoxon test, $p < .02$; Figure 2.5B). Additionally, N→A/U errors occurred at a significantly higher rate than do N→G/C errors (Figure 2.5C, Wilcoxon test, $p < .02$). In contrast, these effects are not due solely to the high frequency of C→U errors: even after the removal of C→U errors (see methods), G/C→N errors remain significantly more frequent than A/U→N errors (Figure 2.5B), and N→A/U errors remain significantly more frequent than N→G/C errors (Figure 2.5C).

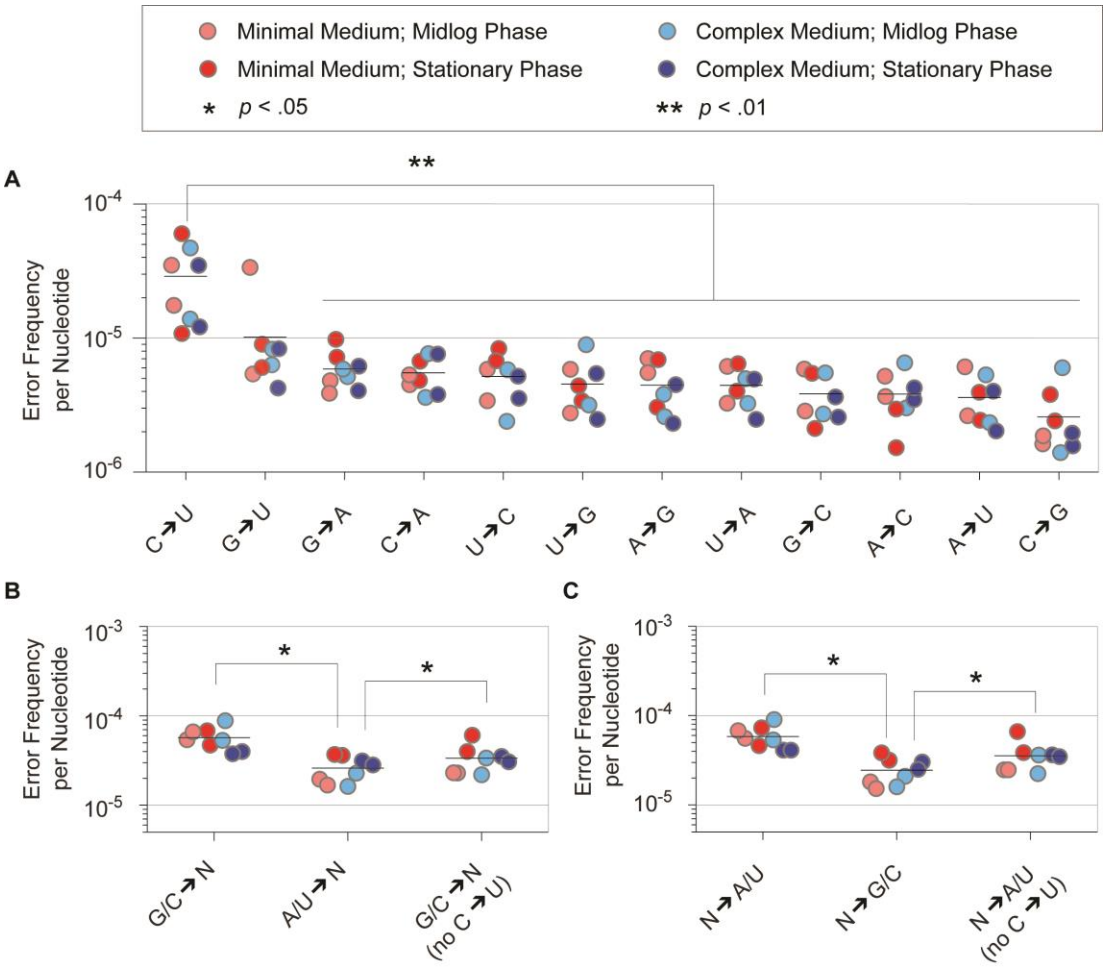

Figure 2.5 – Transcription error frequencies by substitution type in *E. coli*

Points are color-coded according to growth condition, and horizontal bars represent mean values for each class of base substitution. **A.** Transcription error frequencies for individual substitutions. C→U is the most common transcription error, displaying a significantly higher error rate than each of the other substitutions. **B.** Effect of base composition (G/C or A/U) on transcription error frequencies. Errors occur at significantly higher frequencies when the original nucleotide is a G or C. Removal of C→U errors from the analysis (right-most column) demonstrates that the significant effect does not depend on the most abundant type of error. **C.** Transcription error frequencies grouped according to base composition G/C or A/U) of resulting substitutions. Transcription errors resulting in A or U occur at significantly higher levels than those resulting in G or C. Removal of C→U errors from the analysis (right-most column) demonstrates that the significant effect does not depend on the most abundant type of error. Comparisons were made by pairwise Wilcoxon tests (n = 8 for each test), subjected to Bonferroni correction: * $p < .05$; ** $p < .01$.

2.3.4 Transcription error rates in host-restricted bacteria with reduced genomes

The bacterial endosymbionts, *Buchnera aphidicola* and *Carsonella ruddii*, harbor small genomes (450 kb and 190 kb, respectively) and have very high substitution rates, as a consequence of both their lack of several repair mechanisms (Figure 2.1) and the reduced efficacy of selection due to their small effective population sizes. These features are also expected to augment rates of transcription errors, so we assayed the transcription error rates in these endosymbionts using methods identical to those employed for *E. coli*. For the replicate samples of *B. aphidicola*, we detected a total of 302 transcription errors in total mRNA, yielding a transcription error rate of 4.67×10^{-5}, which is significantly different than the rate that we obtained for *E. coli* mRNA (two-tailed *t*-test, $t(8) = 2.28$, $p = .04$) (Figure 2.6A).

Transcription errors in *C. ruddii* mRNA could not be assigned unequivocally because the *C. ruddii* RNA was extracted from a natural population of individuals, rendering it difficult to distinguish between transcription errors and the polymorphisms that might be present in the population. Instead, we quantified transcription error rates for 16S and 23S ribosomal RNA in both *C. ruddii* and *B. aphidicola* because these operons are present in single copy, have high read-coverage (despite the rRNA removal step) and are not polymorphic within a species. (The multiple polymorphic rRNA operons within the *E. coli* genome make estimating rRNA transcription error rates unfeasible in *E. coli*.) We detected a total of 1,229 errors in *C. ruddii* rRNAs and 6,777 errors in *B. aphidicola* rRNAs, yielding rRNA transcription error rates of 5.09×10^{-5} for *C. ruddii* and 3.28×10^{-5} for *B. aphidicola* (Figure 2.6A). Our estimates of bacterial transcription error rates

are, in descending order, 8.23 x 10^{-5} for *E. coli* mRNA, 5.09 x 10^{-5} for *C. ruddii* rRNA, 4.67 x 10^{-5} for *Buchnera* mRNA, and 3.28 x 10^{-5} for *Buchnera* rRNA. The transcription error rates for *B. aphidicola* mRNA and rRNA do not differ significantly from one another.

2.3.5 Biases in endosymbiont transcription error rates

Assessing the transcription errors occurring in both *Buchnera* mRNA and rRNA allowed us to determine whether there are any observable differences in the error rates for two RNA substrates, as might be caused by base compositional biases or selection. All possible nucleotide substitutions, as attributable to transcription errors, were detected in both the mRNA and rRNA samples (although one of *B. aphidicola* mRNA replicate lacked any A→C changes). There were no significant differences for any of the individual substitution classes between mRNA and rRNA, or among any of individual substitution classes (Figure 2.6B).

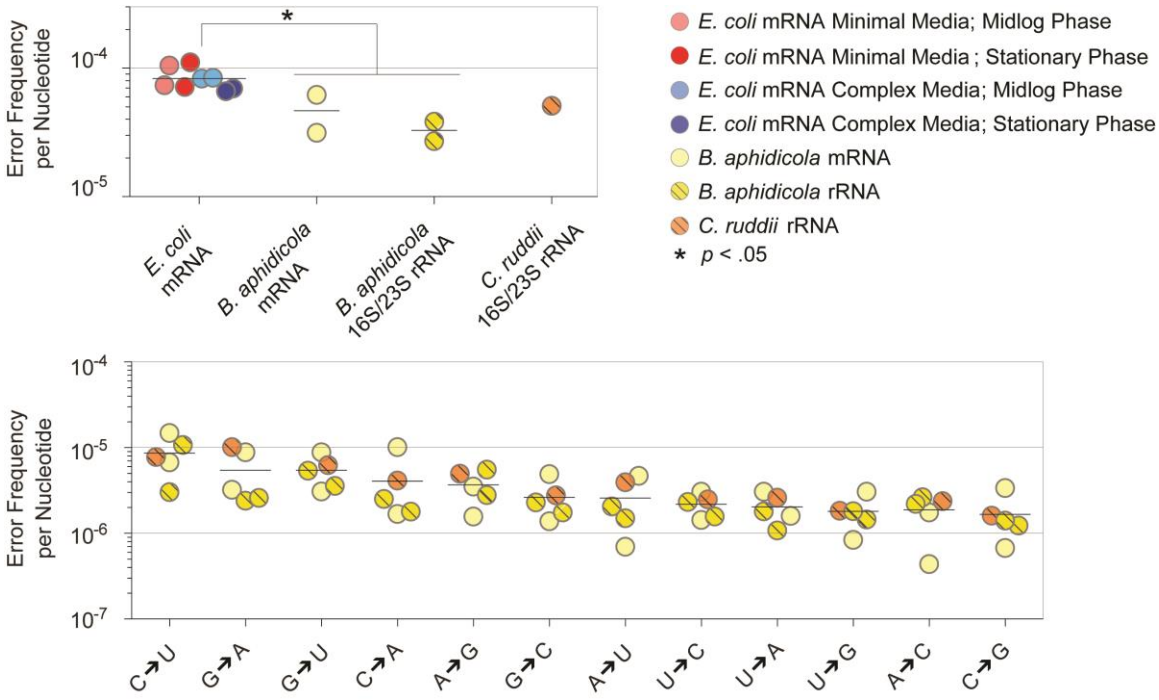

Figure 2.6 – Transcription error frequencies in divergent bacterial taxa

Points are color-coded according to growth condition or source of RNA (see Key), and horizontal bars represent means of each column. **A.** Transcription error frequencies in *E. coli* mRNA (n = 8), *B. aphidicola* mRNA (n = 2), *B. aphidicola* rRNA (n = 2), and *C. ruddii* rRNA (n = 1). **B.** Transcription error frequencies by substitution type in bacterial endosymbionts. No significant differences were detected for any pairwise comparisons.

2.3.6 Effects of transcription errors on protein sequences

Given that each transcript can be translated—perhaps multiple times—into protein, we determined which transcription errors result in an amino acid substitution. On average, 68 of transcription errors cause an amino acid substitution in *E. coli*, whereas 80% of the transcription errors in *Buchnera* result in amino acid substitutions. If errors were to occur at random over the *E. coli* transcriptome, the probability of changing an

amino acid is significantly higher than that actually incurred by transcription errors (76% *vs*. 68%; pairwise Wilcoxon test, $p < .008$).

2.4 DISCUSSION

Considering the range of variation in replication and translation error rates both within and among bacterial species, our finding that transcription error rates are similar for different species, for different classes of RNA sequences, and under different physiological conditions within a species, is bewildering. The mutation (*i.e.*, DNA replication error) rates for bacteria span by several orders of magnitude (94); and for the specific organisms that we consider, spontaneous mutation rates vary nearly 50-fold, from 8.9×10^{-11} per site per generation in *E. coli* (94) to 4.0×10^{-9} for *Buchnera aphidicola* (105). In contrast, based on our genome-wide deep-sequencing approach, the transcription error rates of these two species differ by less than two-fold (4.67×10^{-5} *vs*. 8.23×10^{-5}), with *E. coli* having the slightly higher rate. Our initial prediction was that endosymbionts would have higher transcription error rates because they are subject to high levels of genetic drift and would therefore sustain more deleterious mutations; however, neither of the studied endosymbionts had elevated transcription error rates.

We reasoned that differential regulation of transcription fidelity factors, such as *greA* (10, 106), *greB* (10), or *dksA* (10, 75), operating during transcription, translation, or protein degradation could provide a mechanism for *E. coli* to modulate its transcription error rate under various conditions and growth phases. The conservation of transcription error rates among species is all the more surprising given that these endosymbionts lack

homologs for several of these transcription fidelity factors (Figure 2.1). Endosymbionts possess the most highly reduced bacterial genomes (107), and the genome sizes of *Buchnera* and *Carsonella* are only 641 kb and 160 kb, respectively (108, 109), in contrast to the 4,640-kb genome of the *E. coli* MG1655. Genome reduction in endosymbionts results from elimination of genes that are no longer necessary in the host environment but also involves the loss of apparently beneficial genes, such as those that enhance the efficiency of universal cellular processes, such as DNA repair, translation, and transcription (Figure 2.1). The lack of certain DNA repair enzymes in endosymbionts have been implicated in their extreme base compositions and increased mutation rates (105, 110, 111); however, loss of multiple RNA fidelity factors, such as *greA* in *Buchnera* (Figure 2.1), and *greA*, *greB,* and *dksA* in *Carsonella* (Figure 2.1), seems not to affect transcription error rates.

These bacterial endosymbionts are missing transcription fidelity factors, but their transcription error rates are unchanged, implying that there are mutations within RNAP that can increase the fidelity of transcription. If there is indeed an optimal transcription error rate across bacteria, selection may have improved the intrinsic error rate in the endosymbiont RNAPs after they lost the transcription fidelity factors. However, neither of the RNAPs of the endosymbionts possess a mutation known to increase transcription fidelity in *E. coli* (47). It is possible that endosymbionts do not require rapid transcription and can tolerate slow but accurate transcription (47). The presence of these fidelity factors in *E. coli* could allow its RNAP to make more errors (which are then corrected),

as a result of selection for increased transcription speeds and increased growth rates (112).

Not only were transcription error rates similar in proteobacterial taxa of vastly different lifestyles, population structures, genomes sizes and mutation rates, but the error rates were comparable across organisms for different broad categories of RNAs. Because structural RNAs (16S and 23 S rRNAs) persist longer than mRNAs, they can incur more damage (due to oxidative stress or deamination), thereby leading to an increase in our estimates of the error rate for ribosomal RNAs. On the other hand, one might anticipate rRNAs to have lower error rates than mRNAs, since sub-functional molecules would be preferentially targeted for degradation (104), leaving only those rRNAs that do not contain errors. It should be noted that under both scenarios the error rate during transcription does not change, but rather the variation in the estimated error rates is caused by differences in the fate of rRNAs after transcription. We were only able to measure transcription error rates for both mRNA and rRNA in *Buchnera*. The average error rate for *Buchnera* mRNA was slightly lower than for rRNA, but this estimate was based on the detection of many fewer errors, and there is no significant difference between the two categories of RNAs (Figure 2.6). It is not possible to measure transcription errors in rRNAs of *E. coli* and in mRNAs in *Carsonella*—in both cases, DNA polymorphisms inherent to the sample prevent recognition of transcription errors.

Unlike what we observe for transcription error rates, the mutation rate of an individual strain can vary depending on its growth conditions. *E. coli* mutation rates have been shown to increase by an order of magnitude during stationary phase and under

nutrient-limited conditions (113). Much of the variation in the mutation rate within a species has been attributed to expression of error-prone polymerases during stationary phase (114, 115) and to increased chemical damage occurring during the switch from exponential growth to stationary growth (116–118). Such chemical damage to DNA is usually corrected through DNA repair pathways, but because analogous pathways do not exist for RNA, it is potentially more susceptible to this source of damage. That there is no increase in either the rate or spectrum of errors to RNA during stationary phase suggests that other mechanisms compensate for stationary-phase stresses (*e.g.*, dps protein and catalases) (119–121) or that RNA is too short-lived to be significantly affected.

The relative frequencies of each type of transcription error were similar across organisms and across growth conditions (Figure 2.5) and correspond to what is observed for spontaneous mutations in these organisms (*i.e.*, that C→U substitutions constitute the most common class of errors, and that A/T→T/A and G/C→C/G transversions occur at some of the lowest frequencies) (96, 107, 122–124). Cytosine is the most unstable nucleobase and has an even higher rate of deamination to uracil when nucleic acids are in a single-stranded state (125), so the pronounced bias towards this error is expected. Therefore, some of the observed transcription errors appear to be due to damage to RNA, although current methods simply enumerate errors and do not discriminate between those caused by base misincorporations occurring during transcription and by damage to the RNA after transcription.

Many of the initial measurements of transcription errors in bacteria were restricted to single reporter genes and assayed the combined effects of transcription and

translation errors by assessing how frequent functional proteins were produced from a mutant gene (4, 5, 12). These assays considered errors in translation to be relatively rare because, in this system, it was thought that only the first ribosome on a transcript would be capable of mistranslation and that most errors could be ascribed to the process of transcription (4, 5). However, translation errors in *E. coli* can occur at rates between 10^{-3} and 10^{-4} per codon (91–93), suggesting that many of the original measurements of transcription errors are confounded by the inclusion of translation errors. Furthermore, the transcription error rates varied by up to an order of magnitude for different stop codons (5), indicating that these fluctuations may be attributed to different translation error rates for different codons (82), therefore the rates derived from these studies require validation by methods that only consider transcription errors.

Previous studies reported that the combined transcription/translation error rate, as inferred from the frequency of errors in protein sequences, increases both in stationary phase and under starvation conditions (8, 91, 93). Because we detected no differences in transcription error rates under these different growth conditions, we reason that this variation manifests during translation and is most likely caused by tRNA scarcity during stationary phase (82, 93, 126). Although decreases in nucleotide concentration also occur during stationary phase (127), this has little effect on the overall fidelity of gene expression. Decreases in nucleotide concentration have been shown to increase the frequency of transcriptional pausing (128), which is closely associated with base misincorporations during transcription (47, 62, 129), so it seems that either (*i*) ribonucleotide concentration does not decrease enough under our experimental conditions

to significantly alter the transcription error rate, or (*ii*) that ribonucleotide concentration-induced pausing does not result from transcription errors. Nonetheless, it is curious that cellular growth conditions modify both the rate of DNA mutations and the rate of protein translation errors but not the transcription error rate.

Rates of translation errors have been estimated as being at least an order of magnitude higher than rates of transcription errors, but because most transcripts are translated multiple times, the realized number of modified proteins originating from transcription errors will equal or exceed the number caused by translation errors. This amplification of individual transcription errors into multiple proteins is likely to account for the reduction of transcription *vs*. translation error rates (10^{-5} *vs*. 10^{-4}).

It has been suggested that errors in proteins, as caused by transcription and translation errors, contribute to survivability in the face of external stresses by the production of novel proteins or metabolites (7, 8, 77) or by inducing the general stress response (84). Such effects could not be accomplished through genomic mutations since such mutations can incur permanent decrements to fitness after the stress is removed. Although transcription errors can increase cellular noise and confer a benefit under certain temporary conditions, most variation introduced by errors will not be advantageous. Thus the predominant direction of selection is to lower error rates since too many errors will overload the proteome with deleterious proteins. Whether or not the above argument is tenable, our findings, showing a remarkable consistency of transcription error rates across ecologically diverse bacterial species, different RNA categories, and under a variety of stress and non-stress growth conditions indicate that

transcription errors would contribute very little to such transient protein errors. Transcription is a much less accurate process than DNA replication, and because transcription errors are not heritable (and the vast majority of RNAs are transcribed faithfully under any set of conditions), there appears to be little selection to modulate the overall transcription error rate.

2.5 METHODS

2.5.1 Strains and growth conditions

Transcription errors were enumerated for *E. coli* MG1655 grown at 37°C in (*i*) 15 g/L tryptic soy broth (TSB) or (*ii*) M9 minimal media supplemented with 0.4% glucose. Bacterial cultures were preconditioned in either TSB or M9 minimal media for 24 hr prior to inoculation for sampling. Overnight cultures were diluted to OD_{600} = 0.05 into fresh media and sampled at mid-log phase (4 hr for TSB; 6 hr for M9) and stationary phase (18 hr for TSB; 24 hr for M9).

Transcription errors were enumerated for *Buchnera aphidicola*, an insect endosymbiont recovered directly from its aphid host, *Acyrthosiphon pisum*. *B. aphidicola* were isolated from five grams of adult aphids by a membrane filtration method (130) as follows: Aphids were crushed by mortar and pestle in 15 ml of Buffer A (25 mM KCl, 35 mM Tris–HCl, 100 mM EDTA, 250 mM sucrose, pH 8.0) at 4°C, and the homogenate was centrifuged at 1,500 x g for 15 min. Pellets were resuspended in 15 ml of Buffer A and passed serially through 100 µM, 20 µM, 8 µM, and 5 µM filters. *B. aphidicola* cells were recovered from the filtrate by centrifugation. Transcription errors occurring in the

genome of *Carsonella ruddii*, another insect endosymbiont, were determined from a pooled sample of bacteriocytes from 200 dissected larvae of the psyllid *Pachypsylla venusta* collected locally from galls present on a hackberry tree. Bacteriocytes were stored in Buffer A at –20°C prior to RNA extraction.

2.5.2 RNA extractions

RNA was extracted from *E. coli* following the RNA*snap* protocol for Gram-negative bacteria (131). Roughly 10^8 bacterial cells were harvested by centrifugation at 16,000 x g for 30 sec, the supernatant was removed by aspiration, and pelleted cells were immediately transferred to liquid nitrogen to halt transcription. Samples were transferred to ice, mixed with 100 µl *of* RNA*snap* solution (18 mM EDTA, 0.025% SDS, 1% 2-mercaptoethanol, 95% formamide), briefly vortexed and incubated for 7 min at 95°C. Following incubation, samples were centrifuged at 16,000 x g for 5 min. Supernatants were mixed with an equal volume of PCI (phenol/chloroform/isoamyl alcohol, 25:24:1), the aqueous phase removed and treated with an equal volume of chloroform, and RNA was precipitated by addition of 1/10 volume 3M sodium acetate, 1/50 volume 50 mg/ml glycogen and three volumes of 100% ethanol. DNA contamination was tested using a Qubit high sensitivity DNA assay (Life Technologies), and RNA quality was assessed on an Agilent Bioanalyzer. Ribosomal RNAs were removed from total RNA preparation using the MICROB*Express* kit (Life Technologies).

RNA was extracted from *B. aphidicola* and *C. ruddii* by the addition of 0.75 ml TRIzol reagent (Life Technologies) to 0.25 ml of harvested cells (or bacteriocytes in the

case of *C. ruddii*). Samples were mixed with 0.5 ml sterile zirconium beads, vortexed for 2 min to disrupt cells and incubated for 5 min at 20°C. Following a chloroform extraction, nucleic acids were precipitated from the aqueous phase by the addition of 1/10 volume 3M Sodium acetate, 1/50 volume 50 mg/ml glycogen, and an equal volume of 100% isopropyl alcohol. Precipitated nucleic acids were washed twice with 70% ethanol, suspended in 50 µl RNase-free ddH_2O and treated with DNase, according to the supplier's specifications (Promega). Reactions were terminated by the addition of an equal volume of PCI, and total RNA was precipitated, quantified, extracted, tested for purity and cleared of ribosomal RNAs as described above.

2.5.3 Library preparation and sequencing

We applied two library preparation procedures that have been reported to differentiate errors that occur during transcription from those that arise during sequencing (13, 14, 89). Both methods aim to produce multiple cDNA copies of each mRNA and identify consensus errors, which represent those that are actually present in the corresponding mRNA template. The first method involves successive rounds of sequencing streptavidin-captured mRNAs (89) to generate the multiple cDNA copies of each mRNA; and the second method (termed 'CirSeq'; (13, 14) is based on the sequencing of short, circularized fragments of mRNA that are copied multiple times by rolling-circle amplification prior to sequencing. Attempts at the original streptavidin-capture method of Gout *et al*. (2013) failed to generate multiple copies of cDNA from each mRNA, and even after consulting with the authors and applying several suggested

additions and modifications to the published protocol, we concluded that this method, as currently described, cannot be used to estimate transcription error rates.

For the CirSeq procedure, we followed the protocol of Acevedo and Andino (2014b) with the following modifications that reduced the total number of steps. Starting with 1 µg of purified mRNA, samples were fragmented with the NEB Magnesium Fragmentation module at 94°C for 5 min and then assayed by denaturing PAGE. Regions of the gel containing RNA fragments in the 80–100 nt size-range were excised from the gel, and RNA was eluted from crushed gel slices by overnight incubation in a solution containing 600 mM sodium acetate, 0.017% wt/vol SDS and 1.67 mM EDTA at 4°C. RNA was recovered from the eluent by ethanol precipitation, washed in 70% EtOH, resuspended in 14 µl ddH_2O and analyzed for quality on an Agilent Bioanalyzer RNA chip. RNA fragments were circularized by incubating the entire sample volume with 1 µl T4 Polynucleotide Kinase (NEB), 1 µl T4 RNA Ligase I (NEB), 2 µl T4 RNA Ligase Buffer (NEB), and 2 µl 10 mM ATP for 30 min at 37°C. Samples were purified by PCI extraction and ethanol precipitation, and libraries were prepared for Illumina sequencing by following the protocol accompanying the NEBNext Ultra RNA Library Prep Kit through completion of the second strand synthesis step. After this step, samples were re-purified by PCI extraction and ethanol precipitation, and analyzed with an Agilent Bioanalyzer RNA chip to determine the extent of rolling circle amplification, which occurred during the cDNA synthesis step of the NEB protocol. After confirming amplification status, ddH_2O was added to a final volume of 200 µl, and samples were subjected to 12 min of pulsed sonication (15 sec on, 15 sec off, amplitude 20%) in a

Qsonica sonicator to obtain fragments for sequencing. After harvesting nucleic acids by EtOH precipitation, we resumed the NEBNext Ultra RNA Library Prep Kit protocol for a target insert size of 300 bp. Samples were barcoded using NEBNext Multiplex Oligos (Index Primers Set 1), and the resulting libraries were sequenced on an Illumina MiSeq using 300 nt reads. Sequencing files were discriminated based on their identifying barcodes and analyzed using the CirSeq_v3 pipeline (13).

2.5.4 Sequence processing and error-rate calculations

Transcription error rates were calculated by recovering all errors in the output file processed by CircSeq_v3. This pipeline is described in detail (13), but briefly, repeats within each read were identified by CircSeq_v3, and aligned to obtain a consensus sequence if a read contained were at least three full repeats of 100 bp or less. Any read that failed to meet this criterion was discarded. Because each base within each repeat is assigned a different quality score, a single quality score representative of the consensus sequence at each base was calculated as the average quality score from the three bases from each repeat at each location. Reads are then mapped to their respective reference genome, and errors were identified as those bases within reads that did not match the reference genome. Only bases that had an average quality score of 20 or higher (see Figure 2.7 and below) were used. Overall per base coverage was calculated as the sum of the total coverage of each base, and overall error rates were calculated by dividing the number of errors by the overall per base coverage. The error rate for each type of

nucleotide substitution, with A→C as an example, was calculated as above except the error rate was adjusted for the base composition of the sequenced RNA such that:

$$r_{adj}^{A\to C} = r^{A\to C} * f_A$$

where $r^{A\to C}$ is the error rate for A→C errors, $r_{adj}^{A\to C}$ is the base adjusted error rate, and f_A is the adjustment coefficient for base composition, calculated as:

$$f_A = \frac{0.25}{A}$$

where A is the fraction of overall adenosine nucleotides sequenced in the transcriptome. This calculation normalizes the error rate of A→C errors by any base compositional biases in the transcriptome. This error rate is presented in the context of the entire transcriptome (*i.e.*, not within the context of all sequenced adenosine locations).

To ensure that sequencing errors did not influence our results, we analyzed the original sequence data to include all bases having an average quality score of 10 and higher, and sequentially increased the stringency of the analysis by analyzing nucleotides at different quality score cut-offs (Figure 2.7). By sequentially increasing the stringency of the analysis, we determined the influence of sequencing errors at each quality score. Transcription error rates asymptote in the quality-score range of 18 to 20 (Figure 2.7), reflecting the point where sequencing errors are removed from the analysis. We selected a quality-score value of 20 for all analyses, a value that maximizes the numbers of actual errors and provides accurate measures of transcription error rates.

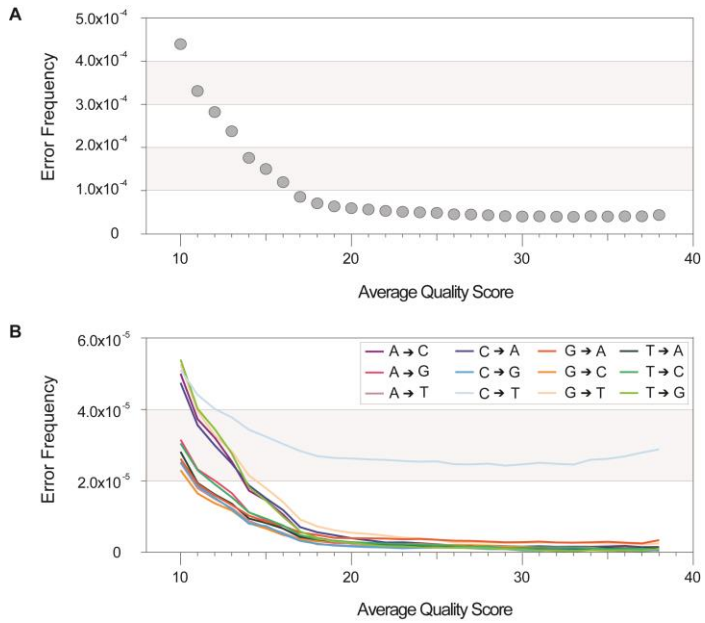

Figure 2.7 – Effect of sequencing errors and data quality on the estimation of transcription error frequencies

Transcription error frequencies for the combined *E. coli* replicates were calculated at increasing average base quality scores between 10 and 40 to demonstrate the effect of sequencing errors and low quality bases on error frequencies. Overall transcription error frequency (**A**) and the transcription error frequency for each nucleotide substitution (**B**) level-off in the quality-score range of 18 to 20, indicating that use of data in this range and beyond excludes sequencing artefacts from estimates of transcription error rates. There were insufficient bases in the transcriptome that attained average quality scores >38 for inclusion in this analysis.

2.5.5 Data processing and analysis

After the sequences were processed by the CirSeq_v3 pipeline with an average quality score cutoff of 20 (Figure 2.7), we removed those duplicate and multicopy genes that are polymorphic within the *E. coli* genome (*e.g.*, structural RNA genes, *ompF* and *ompC*, and *tufA and tufB*) since the source of variation can not be unequivocally assigned.

Transcription error rates were adjusted for base composition of the sample using the weighted average of the occurrence of each nucleotide in the particular individual transcriptome being considered.

We developed custom Python scripts to determine: (*i*) transcription errors, calculated by tabulating the total number of errors identified by the CirSeq_v3 pipeline within the protein coding regions of the genome; (*ii*) nucleotide coverage, calculated by adding the overall coverage of each base within the protein coding regions of the genome; (*iii*) error rates, calculated by tabulating the total number of errors and base coverage of all coding regions within 50-kb non-overlapping windows across an entire genome and dividing the number of errors by the coverage, yielding an error rate; (*iv*) leading/lagging strand error rates, calculated by tabulating the errors and coverage of all genes situated on either the leading or lagging strands and calculating the error rate as above; (*v*) the number of errors that would result in an amino acid replacement by chance, calculated by randomly generating simulated transcription errors from each sequenced transcriptome and determining their effects on the amino acid sequence. All statistics were performed in Prism Graphpad or R.

The list of nucleic acid information processing genes and the associated functions were curated using EcoCyc (132). Orthologs of these genes in the endosymbionts were determined using BLASTP from NCBI (http://blast.ncbi.nlm.nih.gov/Blast.cgi) with an E score cutoff of ≤1 and an amino acid positive score cutoff of ≥40%. The genome accession number is NZ_ACFK01000001 for *B. aphidicola* LSR1 and NC_008512 for *C. ruddii* PV.

Chapter 3: Genome-wide spectra of transcription insertions and deletions reveal that slippage depends on RNA:DNA hybrid complementarity[1]

3.1 ABSTRACT

Advances in sequencing technologies have enabled direct quantification of genome-wide errors that occur during RNA transcription. These errors occur at rates that are orders of magnitude higher than during DNA replication, but due to technical difficulties, such measurements have been limited to single-base substitutions and have not yet quantified the scope of transcription insertions and deletions. Previous reporter-gene assays suggest that transcription indels are produced exclusively by elongation-complex slippage at homopolymeric runs, so we enumerated indels across the protein-coding transcriptomes of *Escherichia coli* and *Buchnera aphidicola*, which differ widely in their genomic base compositions and incidence of repeat regions. As anticipated from prior assays, transcription insertions prevail in homopolymeric runs of A and T; however, transcription deletions arise in much more complex sequences and are rarely associated with homopolymeric runs. By reconstructing the relocated positions of the elongation complex as inferred from the sequences inserted or deleted during transcription, we show that continuation of transcription after slippage hinges on the degree of nucleotide complementarity within the RNA:DNA hybrid at the new DNA template location.

[1] This chapter is reproduced (with minor modifications) from its initial publication:

Traverse CC and Ochman H (2017) Genome-wide spectra of transcription indels reveal that slippage depends on RNA:DNA hybrid complementarity. *MBio* 8:e01230-17. Ochman H supervised the project.

3.2 INTRODUCTION

In addition to the errors that occur during DNA synthesis, which form the basis for adaptation and heritable genetic variation, non-heritable errors are generated during the process of transcription. These transient errors are produced at rates that are orders of magnitude higher than replication error rates (2–4, 133), such that the cell will invariably express a subset of transcripts that do not match the encoded sequence. This non-heritable variation is most often considered to be deleterious since it can burden the cell with faulty or misfolded proteins in a similar manner to DNA mutations. Transcription errors can also result in collisions between replication and transcription machineries, thereby generating double-strand breaks in the chromosome and abortion of the transcript (36, 37, 134, 135). It has also been proposed that transcription errors might somehow provide beneficial variation during times of stress (7, 8, 84, 85). Nonetheless, because transcription errors are not stable across generations, studying their incidence and patterns of occurrence has traditionally been difficult.

Transcription error rates were originally measured with reporter genes engineered with premature stop codons, such that a specific transcription error would convert the sequence to produce a functional reporter protein (4, 136). Recently, a high-throughput sequencing approach expanded the spectrum of transcription errors from assaying a single site in a reporter gene to all protein-coding nucleotides in the transcriptome (13, 14). This technique, which relies on a unique library preparation method, allows for direct quantification of transcription errors without contamination by the sequencing

errors that typically befall RNAseq methodologies. Results from a study applying this method revealed a transcription base-substitution rate in *Escherichia coli* of ~8 x 10^{-5} per nucleotide that was relatively constant across different growth states and growth phases (3). Whereas this approach can provide accurate, genome-wide measurements of transcription errors, all sequencing-based studies in bacteria have been confined to the detection of base substitutions and have ignored transcription insertions and deletions (indels) (2, 3, 90, 133).

Transcription indels may be more detrimental than base substitutions because individual nucleotide changes only alter a single amino acid and are often silent, whereas indels can involve multiple amino acids and will usually disrupt the reading frame. The indels generated during transcription are generally thought to occur through forward- or backward-slippage of the actively transcribing RNA polymerase (elongation complex) along the template DNA, causing a portion of the template to be either skipped (resulting in a deletion) or re-transcribed (leading to an insertion) (9, 12, 137, 138). Previous work exploited this slippage mechanism as a way of detecting transcription indels by engineering reporter genes with homopolymeric runs that, upon slippage, restored the proper reading frame (9, 12, 137, 138). Although these studies yielded information about relative indel rates in certain homopolymeric tracts, such repeats are inherently error-prone and are not likely to represent the indel rate in coding sequences, which only rarely contain long homopolymeric runs (139).

The focus on error-prone repeats led to the notion that transcription indels occur primarily at homopolymeric runs (9, 12, 137, 138), which are being selectively removed from genomes (140). In this chapter, we evaluate the occurrence of transcription indels throughout the entire genome, and gain insights into the substrates and mechanism of transcriptional slippage. Because most information concerning transcription slippage has focused on homopolymeric runs, we compare the rates and patterns of indels in the transcriptomes of *E. coli*, which has an equitable occurrence of each nucleotide, and a low G+C bacterial endosymbiont whose genome is greatly enriched in long tracts of adenosine and thymine. We found that while insertions predominate in homopolymeric runs in both species, deletions occur in more complex sequences. These results led us to develop a general model of transcription slippage that is driven by RNA:DNA hybrid complementarity at the site of the new DNA template.

3.3 RESULTS

3.3.1 Rates of transcription-induced indels across the transcriptome

We analyzed the spectrum of insertion and deletion errors across all coding regions of the transcriptomes of *Escherichia coli* and *Buchnera aphidicola* by applying a circularization method that prevents the inclusion of sequencing artefacts (14). For the eight replicate samples of *E. coli*, transcription errors causing deletions vastly outnumbered those causing insertions, 921 to 72, yielding an average rate of 1.57 x 10^{-5} deletion events and 1.35 x 10^{-6} insertion events per transcribed nucleotide (Figure 3.1). In *Buchnera*, however, the preponderance of transcription indels were insertions:

across the two replicates, there was a total of 157 insertions and 70 deletions, representing 1.30 x 10^{-5} insertion events and 1.75 x 10^{-5} deletion events per transcribed nucleotide (The mean insertion rate is higher than the mean deletion rate in *Buchnera* due to the high variance among samples; Figure 3.1). Despite their contrasting patterns, the overall rates of transcription indels in *E. coli* and *Buchnera* differ by less than two-fold; and in *E. coli*, the overall rate of transcription indels is within the same order of magnitude as transcription errors that result in base substitutions (Figure 3.1). Considering both nucleotide substitutions and indels, the cumulative transcription error rate per transcribed nucleotide is 9.94 x 10^{-5} in *E. coli* and 7.73 x 10^{-5} in *Buchnera*. We found no effects of transcript expression level, gene orientation, or error location within a transcript on the transcription indel rates in both *E. coli* and *Buchnera*.

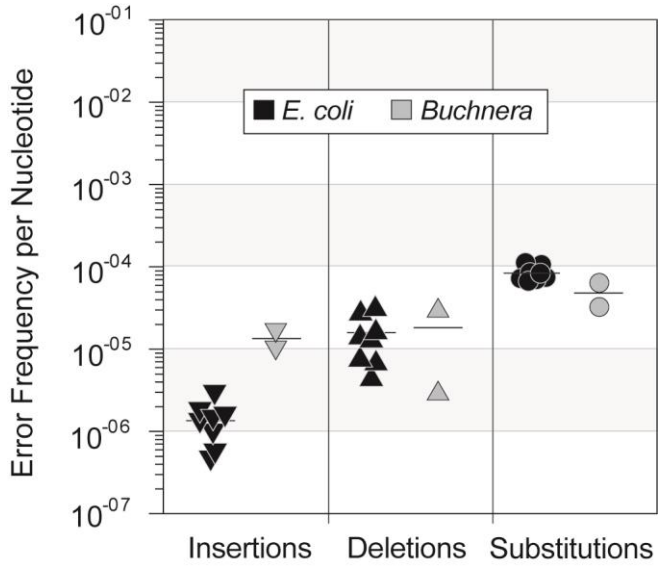

Figure 3.1 – Rates of transcription insertions, deletions, and base substitutions in *E. coli* and *Buchnera*.

Frequencies of each type of transcription error were computed for eight replicate samples in *E. coli* and for the same two replicate samples in *Buchnera*.

3.3.2 Insertion errors in homopolymeric runs

Among the insertions that occur during transcription, 80% involve the addition of an individual nucleotide (Figure 3.2). These single-nucleotide insertions predominate in homopolymeric runs, and their frequencies increase exponentially with the length of the homopolymeric run (up to the maximum of nine nucleotides in *E. coli* and 12 nucleotides in *Buchnera;* Fig 3.3). In every case, the inserted nucleotide matches those comprising the repeat, suggesting that these errors arise through a backward-slippage mechanism.

The ten-fold difference in the numbers of transcription insertions in *E. coli* and *Buchnera* can be ascribed almost entirely to the incidence of homopolymeric runs in

these genomes. The low (26%) G+C content of the *Buchnera* genome increases the likelihood and lengths of homopolymeric runs of adenine or thymine, which comprise 80% of the sequenced homopolymeric runs in the *E. coli* transcriptome and 97% of the sequenced homopolymeric runs in the *Buchnera* transcriptome. In both organisms, 100% of insertions within homopolymeric runs occurred in runs of adenine or thymine, indicating that homopolymeric runs of guanine and cytosine are not slippage-prone. A minority of transcription insertions (19 in *E. coli* and 6 in *Buchnera*) do not occur in these repeat tracts, but in 15 of the 19 such cases in *E. coli*, the inserted nucleotide(s) match the preceding nucleotides, suggesting that they originated by backward-slippage followed by re-transcription of the slipped region (Table 3.1).

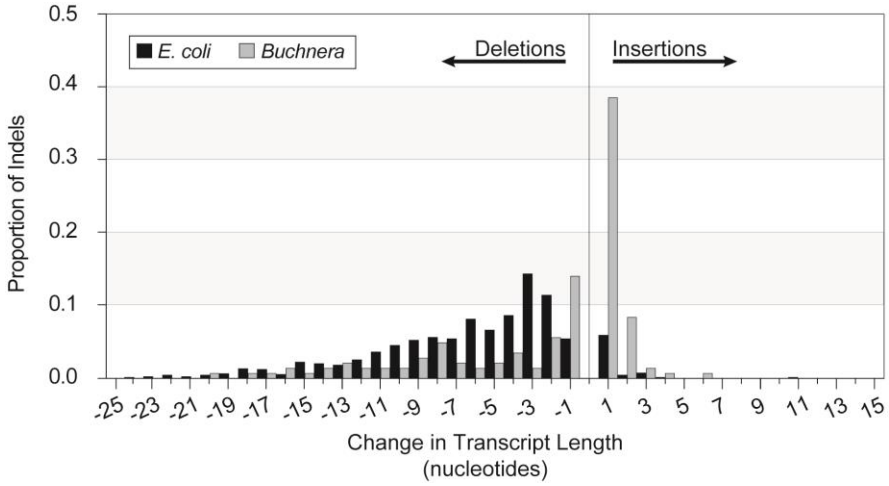

Figure 3.2 – Length distribution of transcription insertions and deletions in *E. coli* and *Buchnera*

Insertions peak at one nucleotide in length but deletions tend to me much longer, peaking around three nucleotides in length in *E. coli*.

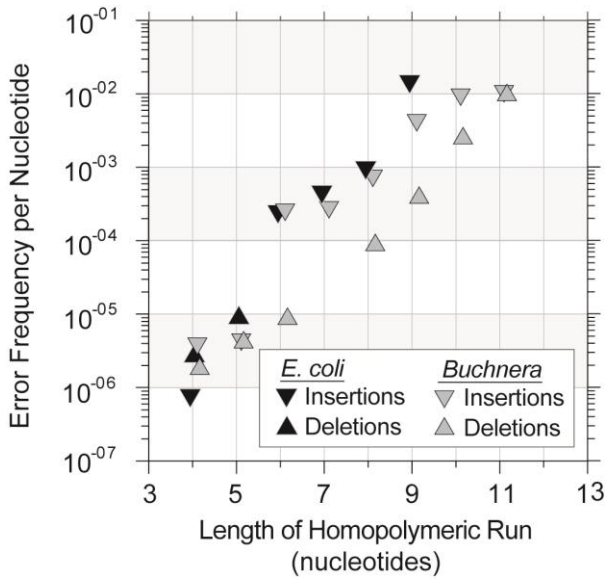

Figure 3.3 – Error frequencies of *Buchnera* transcription insertions, *Buchnera* transcription deletions, and *E. coli* transcription insertions in homopolymeric runs

Each error type shown follows a natural exponential function: $Buchnera_{\text{insertions}}$, $r^2 = .739$, $p < .004$; $Buchnera_{\text{deletions}}$, $r^2 = .981$, $p < .001$, $E.\ coli_{\text{insertions}}$, $r^2 = .894$, $p < .003$. There were too few *E. coli* transcription deletions to test for this trend.

Table 3.1 – *E. coli* transcription insertions in non-homopolymeric region

Preceding Nucleotides[a,b]	Inserted Nucleotides[b]	Succeeding Nucleotides[a]	Insertion Length (Nucleotides)
CGCTGGCGC	**GCCGCTGGCGC**[c]	AATGGATAG	11
T**ATTTATTT**	**ATTT**	CGCCCTGCC	4
GTG**ATGATG**	**ATG**	TATAACCGG	3
AGAAGAAGA	**AGA**	TAAAAACAG	3
TTCTTCTTC	**TTC**	GCGAAGCGT	3
CTC**TTCTTC**	**TTC**	CAGCGTCGG	3
CTTGAG**CCG**	**CCG**	TCGTCGTGG	3
TTCTTCTTC	**TTC**	AACACCGAC	3
AACAACAAC	**AAC**	CGATGAACT	3
CGGTCTGGA	AG	CAAAGGCAC	2
CGGCGGTT**A**	**A**	TTTTTTTGC	1
TCGAAGAA**C**	**C**	GCGTTAAGA	1
TCCGTTCT**A**	**A**	CAAACATTT	1
GAACAGGC**G**	**G**	AAAAAAGTG	1
CTGAAAG**AA**	**A**	GCGGCAGAA	1
TTCGTAG**AA**	**A**	GCTGAGTAA	1
CATACCACC	T	ATCGTTAAG	1
CTGGCAGAA	G	ACGTTATCC	1
ACTGGCGGC	A	GCAAACCGG	1

[a]Columns list the nine preceding and the nine succeeding nucleotides because this number corresponds to the lengths of RNA:DNA hybrids in the elongation complex.

[b]Bold sequences represent instances of slippage followed by re-transcription of the slipped region.

[c]In this case, the inserted nucleotides match 11 of the preceding nucleotides, but only nine are listed in the Preceding Nucleotides column

3.3.3 Transcription deletions in *E. coli* often preserve the reading frame

In contrast to transcription insertions, the majority of transcription deletions entail multiple nucleotides. The spectra of transcription deletions differ in *E. coli* and *Buchnera*, likely stemming from their differences in nucleotide composition. In *E. coli*, trinucleotide deletions, which keep the protein in frame, are more frequent than either mono- or di-nucleotide deletions (Figure 3.2), and deletions within homopolymeric tracts are rare. Additionally, there is a three-nucleotide periodicity in short deletions of six or fewer bases, with peaks at three and six nucleotides in length ($p < .01$, Fisher's exact test). In *Buchnera*, the most common transcription deletion involves single nucleotides (Figure 3.2), over half of which occur in homopolymeric tracts; but only 16 of the 57 deletions in *Buchnera* occurred within homopolymeric tracts as opposed to 151 of the 157 insertions. The error rate of *Buchnera* deletions in homopolymeric runs increases exponentially as the length of the run increases, similar to what was observed for insertions (Fig 3.3).

3.3.4 Transcription deletions are A+U biased

Within *E. coli*, there is a bias in the composition of nucleotides removed by transcription deletions. The average composition of deleted nucleotides is 39.5% G+C, differing significantly from the overall nucleotide composition of 53.3% G+C for coding regions of the transcriptome (pairwise Wilcoxon test, $p < .001$). Moreover, guanine and cytosine are significantly underrepresented within transcription deletions (Figure 3.4A), indicating that certain nucleotide-enriched regions are resistant to slippage. Unlike *E. coli*, the nucleotide contents of transcription deletions in *Buchnera* did not differ

significantly from that of the entire transcriptome (Figure 3.4B), perhaps due to the already elevated A+U content of the *Buchnera* genome.

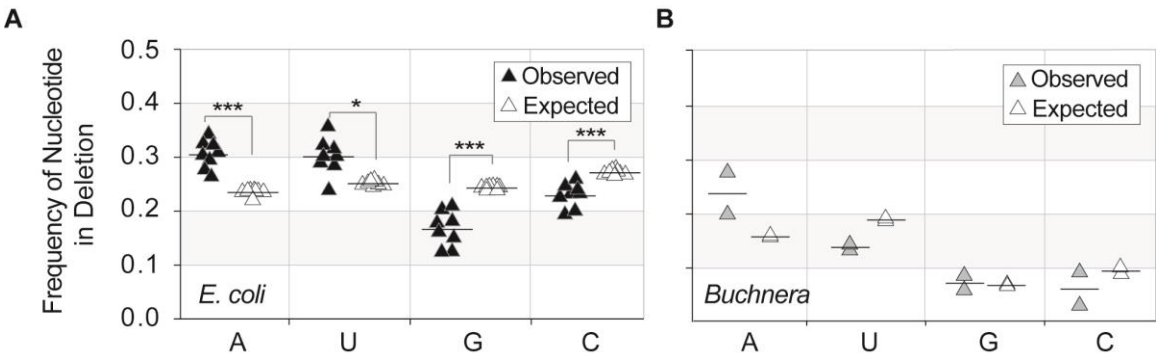

Figure 3.4 – Compositional biases of transcription deletions

A. Nucleotide composition of transcription deletions in *E. coli* (black) compared to that expected based on the nucleotide composition of all transcribed sequences (white) in each replicate. Comparisons were performed using pairwise Wilcoxon tests, subjected to the Benjamini-Hochberg correction (* $p < .05$, *** $p < .001$). **B**. Nucleotide composition of transcription deletions in *Buchnera* (gray) compared to that expected based on the nucleotide composition of all transcribed sequences (white) in each replicate. Comparisons were performed by a Student's t-test

3.3.5 Effects of preceding and succeeding nucleotides on transcription deletions

Because backward slippage, as provoked by certain upstream nucleotides, was found to be a major source of transcription insertions, we asked if there were any nucleotide-compositional biases in the regions preceding and succeeding each deletion (Figure 3.5). The −1 positions, *i.e.* the last nucleotide transcribed before slippage, were significantly enriched in adenines and deficient in guanines. The −2 positions of regions preceding deletions were significantly enriched in cytosine and were again deficient in

guanine, and the −3 and −4 positions had significantly lower cytosine compositions. Additionally, the +1 position, *i.e.* the first nucleotide transcribed after a deletion, was significantly enriched in adenine and had significantly lower guanine composition.

The nucleotide composition of dinucleotides surrounding transcription indels was also biased: Transcription deletions in *E. coli* were more likely to occur immediately after transcription of CA, TA or AT dinucleotides (Fisher's Exact Test; $p < .001$), whereas many of the G-rich dinucleotides (GG, TG, GA, AG, and CG) were less likely to promote a deletion ($p < .001$) (Figure 3.6A). The dinucleotide composition of regions further upstream did not impose any detectable effect on the occurrence of transcription deletions. Although there were insufficient deletions in *Buchnera* to test the influence all dinucleotide pairs, transcription deletions occurred at significantly higher frequencies when CA or AA is the preceding dinucleotide (Figure 3.6B). All significantly higher or lower incidences of trinucleotides could be explained by the trends observed for dinucleotides.

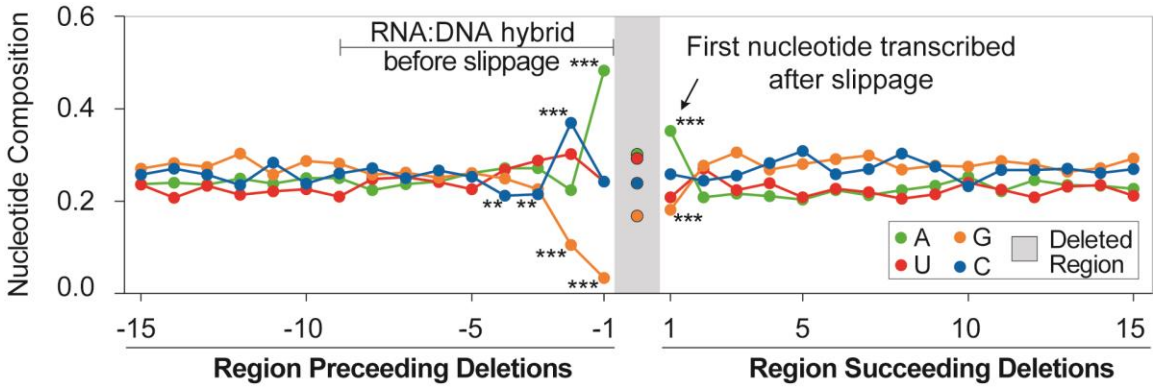

Figure 3.5 – Composition of preceding and succeeding nucleotides relative to deletions

Average proportion of each nucleotide at each of the 15 bases preceding and the 15 bases succeeding transcription deletions, and in deletions as a whole (shaded gray region). Significant biases in nucleotide frequencies occur in the four bases before a deletion and one base after a deletion. Comparisons were made using Fisher's exact test, subjected to the Benjamini-Hochberg correction (* $p < .05$, *** $p < .001$)

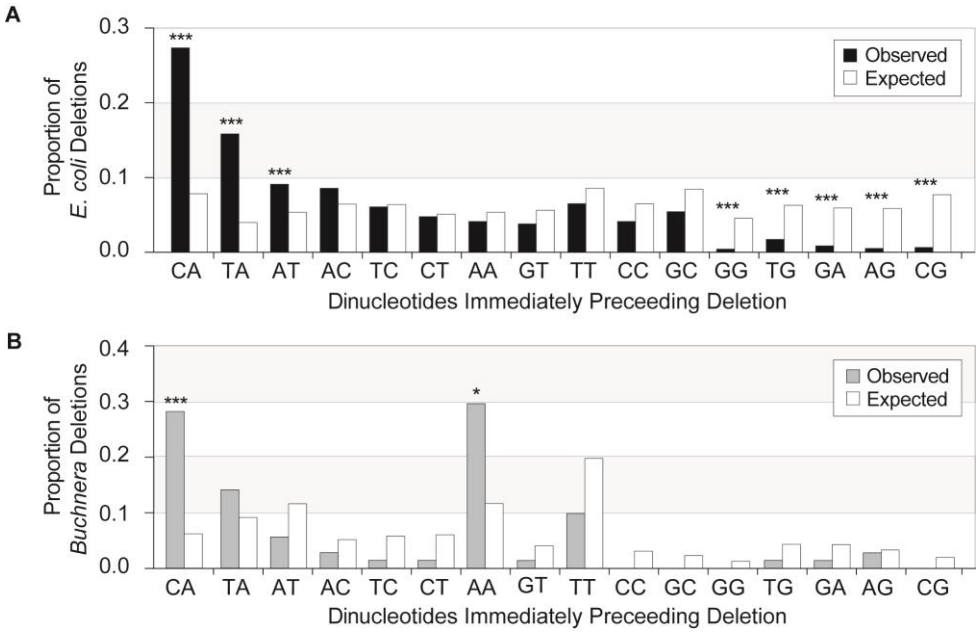

Figure 3.6 – Dinucleotide frequencies of the two bases preceding transcription deletions

A. *E. coli* (black) is compared to that expected based on the nucleotide composition of all transcribed sequences (white). **B**. Dinucleotide frequencies of the two bases preceding transcription deletions in *Buchnera* (gray) compared to that expected based on the nucleotide composition of all transcribed sequences (white). Comparisons were made using Fisher's exact test, subjected to the Benjamini-Hochberg correction (* $p < .05$, *** $p < .001$).

3.3.6 Slippage stops at locations with high RNA:DNA hybrid complementarity

In the current model for transcription deletions, the elongation complex and transcript lose register with the template DNA (such that the transcript and DNA template are no longer paired), slip forward, and then resume transcription at a downstream point on the template DNA (12). After a slippage event, the RNA:DNA hybrid between the

nine most recently transcribed nucleotides and the DNA template commonly contains several mismatches because the elongation complex resides in a new location.

To calculate the extent of complementarity between the slipped transcript and the new DNA template location, we reconstructed the RNA:DNA hybrids after slippage by comparing the nine nucleotides immediately preceding the start of each deletion with the nine nucleotides preceding the end of each deletion, each from the annotated start and end points of the deletion in the reference sequence. In both *E. coli* and *Buchnera*, the reconstructed RNA:DNA hybrids from observed deletions had more complementary base-pairing than expected (Figure 3.7A and 3.7B; Chi-square tests, $p < .0001$), indicating that after a slippage event, transcription is more likely to resume in regions that impart high complementarity within the new RNA:DNA hybrid.

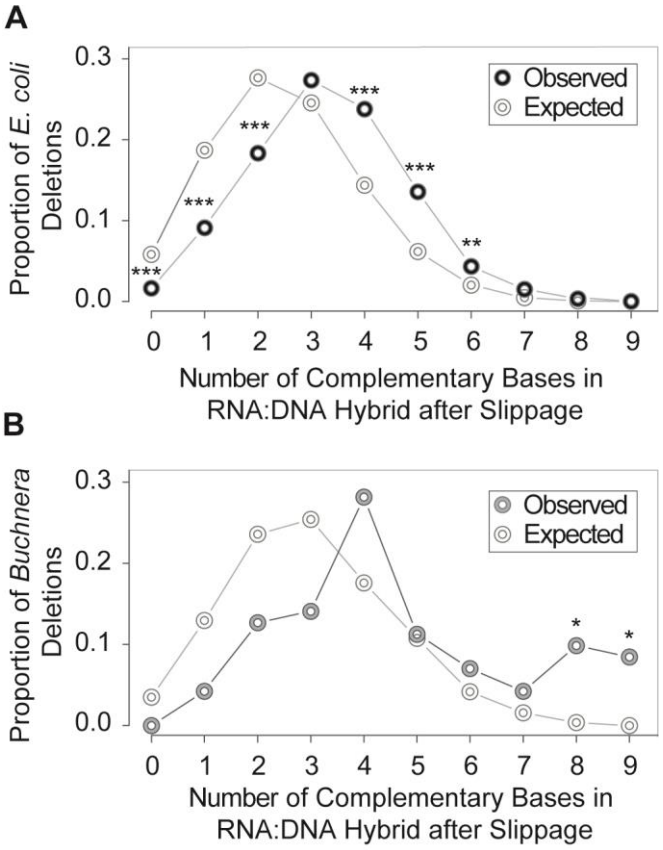

Figure 3.7 – Dependence of transcription deletions on sequence complementarity in the RNA:DNA hybrid.

A. The nine-base RNA:DNA hybrids were reconstructed for transcription deletions (black rings) and for expected deletions based on the nucleotide composition of all transcribed sequences (white rings) in *E. coli*, and the extent of complementarity between the region preceding the end of a deletion and the RNA:DNA hybrids was computed. **B.** The nine-base RNA:DNA hybrids were reconstructed for transcription deletions (grey rings) and for deletions expected based on the nucleotide composition of all transcribed sequences (white rings) in *Buchnera*, and the extent of complementarity between the region preceding the end of a deletion and the RNA:DNA hybrids was computed. For both organisms, there were significant deviations from expectation (Chi-square test, $p < .001$), indicating that transcription slippage is more likely to stop at regions of higher base complementarity than expected. Comparisons of the extent of RNA:DNA complementarity were performed using Fisher's exact test, subjected to the Benjamini-Hochberg correction (* $p < .05$, ** $p < .01$, *** $p < .001$).

3.3.7 Transcriptional deletions are associated with sequence repeats in *E. coli*

We next examined the extent to which non-homopolymeric repeat sequences were associated with transcriptional deletions. Deletions of two or three nucleotides were significantly more likely to occur in regions containing di- or tri-nucleotide repeats, respectively (Figure 3.8A; Wilcoxon test, $p < .01$). Overall, about half of all two-nucleotide deletions occurred in a dinucleotide repeat, and 37 of the 143 three-nucleotide deletions occurred in a trinucleotide repeat. Unlike insertions, these short deletions do not increase in frequency with repeat number: 34 of the 37 repeating runs that promoted deletions consisted of only two repeats, with the second instance of the repeat experiencing the deletion.

To determine if deletions were more likely to occur when repeats are separated by intervening sequences, we enumerated the deletions that were complementary to the new DNA template at the 3'-end of the RNA portion of the RNA:DNA hybrid. Nearly 30% (n = 265) of all deletions had complementary base-pairing in the last two positions in the RNA:DNA hybrid, significantly higher than expected by chance (Figure 3.8B). Additionally, there was an increased occurrence of complementary base-pairing of all nucleotides within the last three and four positions of the slipped transcript and the new DNA template (Wilcoxon test; $p < .01$) (Figure 3.8B). In sum, the final two, three, or four positions in the RNA:DNA hybrids are significantly more likely to experience complementary base-pairing after forward transcription slippage.

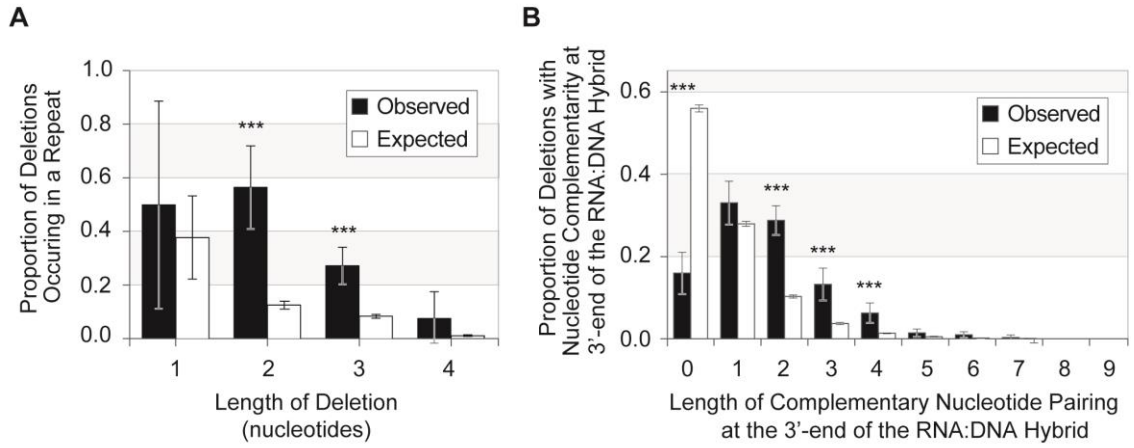

Figure 3.8 – Transcription deletions in short sequence repeats.

A. Proportions of transcription deletions between 1 and 4 nucleotides in length occurring within repetitive sequences in *E. coli*. In all cases, deletion lengths correspond to the length of the repeat unit within a repetitive sequence, and there is a minimum of two repeat units for a sequence to be considered repetitive. (The wide error bars in single nucleotide deletions results from replicates with few or no deletions of that length.) **B.** Proportions of deletions with successive complementary bases in the 3'-end of RNA:DNA hybrid after slippage. Deletions of all lengths included in this analysis. Comparisons in **A** and **B** were performed with pairwise Wilcoxon tests (n = 8 for each test), subjected to the Benjamini-Hochberg correction (* $p < .05$, ** $p < .01$, *** $p < .001$).

3.4 DISCUSSION

Previous assays of the insertions and deletions that arise during transcription examined only those errors occurring within long synthetic homopolymeric repeats and provided neither the absolute rate of transcription indels nor the full spectrum of sequence motifs prone to such errors (9, 12, 137, 138). These issues can now be resolved through the application of a genome-wide approach that assays errors incurred over the

entire transcriptome and furnishes accurate estimates of error rates based on the actual number of nucleotides transcribed.

Cumulatively, there were 993 indels (921 deletions and 72 insertions) in *E. coli* and 227 indels (70 deletions and 157 insertions) in *Buchnera*, yielding rates of 1.7×10^{-5} and 3.1×10^{-5} indels per transcribed nucleotide, respectively. The transcription error rates for indels are several-fold lower, but within the same order of magnitude, as the transcription error rate for base substitutions reported for *E. coli* and *Buchnera* (3), yielding overall transcription error rates of nearly 10^{-4} per transcribed nucleotide. Given an average gene length in bacteria of 10^3 bp denotes that 1 in 10 transcripts suffer some type of transcription error. Such high rates can be tolerated because transcription errors, in contrast to replication errors, are ephemeral and usually affect only a very small fraction of the proteins produced from a given locus; and additionally, there are mechanisms to refold and remove damaged proteins (84, 141, 142). The strong mutational bias of *Buchnera* towards A+T promotes the occurrence of long homopolymeric tracts, which experience frequent indels during replication, giving rise to pseudogenes (105, 143). It has been proposed that transcription indels might serve to correct these frameshifted pseudogenes (143); however, we detected no cases where a transcription error restored a reading frame.

Based on early models of transcription slippage, previous assays of transcription indels were designed to detect errors occurring in homopolymeric runs. With respect to transcription insertions, our results largely corroborate prior findings because our genome-wide approach showed that in both *E. coli* and *Buchnera*, a majority of

transcription insertions involve the addition of a single base into homopolymeric runs of either A or T (9, 12, 137, 138). Those few insertions (18 in *E. coli*, 3 in *Buchnera*) that occurred in non-homopolymeric sequences—errors that were previously never assayed—mostly involved duplications of the preceding nucleotide, suggesting that virtually all transcription insertions, whether in homopolymers or not, are caused by re-transcription after events of backward-slippage.

In contrast to transcription insertions, most transcription deletions occur in sequences that are more complex and were therefore missed by previous assays, which focused solely on transcription errors in homopolymeric runs. Only 15 (21%) transcription deletions in *Buchnera* were initiated within uninterrupted homopolymeric runs, and only 3 (less than 1%) transcription deletions in *E. coli* initiated within uninterrupted homopolymeric runs—a difference likely attributable to the high incidence of homopolymeric runs in *Buchnera*.

Although the relatively high frequencies of transcription errors when compared to replication errors imply that transcription errors are generally of little consequence to cellular fitness, we detected a 3-nt periodicity in deletions of six or fewer nucleotides in *E. coli*, suggesting that selection serves to avoid or eliminate frame-shifting deletions or that the triplet nature of codons imparts higher complementarity in intervals of 3. Transcription deletions are common in *E. coli*, and these ≤6 nt deletions comprise 58% of all deletions in *E. coli*, so it possible that they occur at frequencies high enough to impact fitness. Despite similarity in the rates of transcription deletions in *E. coli* and *Buchnera*,

the periodicity in transcription deletions was not apparent in *Buchnera*, most likely because selection is less effective due to their small effective population sizes.

Knowledge of the full scope of transcription errors provides several insights into the mechanisms by which transcription indels arise. In brief, most deletions were greater than one nucleotide in length, whereas most insertions are one nucleotide in length, arising from the backward-slippage of the elongation complex by only one base before transcription resumes (Figure 3.2). Because multiple elongation complexes can transcribe genes in arrays (144, 145), it is possible that once the nascent transcript loses register with the DNA template and the elongation complex slips backwards, upstream elongation complexes push the slipped elongation complex forward, thereby limiting how far back it can slip. Since the vast majority of deletions are greater than one nucleotide in length (Figure 3.2), following this scenario, it appears that the upstream elongation complexes can propel the slipping elongation complex, causing it to skip forward several nucleotides.

A process by which upstream arrays of elongation complexes *(i)* prevent large insertions by blocking further backward-slippage, *(ii)* help restore the original position of a slipped elongation complex, and *(iii)* facilitate translocation to distant positions, all help explain the low insertion-to-deletion rate in *E. coli*. Once an elongation complex and transcript loses register with the template and slip backward, the forward translocation from an array of actively transcribing elongation complex will most likely result in a deletion rather than an insertion. This process likely operates in *Buchnera* as well, but the

high incidence of homopolymeric runs makes for many more backward-slippage events, thereby elevating the number of insertions.

Although the majority of the transcription insertions originate in homopolymeric runs, several insertions occurred outside of these sequences, suggesting that several mechanistically similar events cause transcription insertions in both *E. coli* and *Buchnera*: *(i)* For insertions at homopolymeric runs, a backward-slip of the elongation complex at these sites usually retains complementary base-pairing between the 3'-end of the transcript and the template DNA, allowing transcription to resume because the template sequence before and after the slippage event remain identical. *(ii)* For those insertions occurring at tri- or tetra-nucleotide repeats, the elongation complex slipped backwards by one repeat and then transcribed an extra repeat (Table 3.1), again retaining complementary base-pairing within a portion of the RNA:DNA hybrid, similar to slippage in homopolymeric regions. *(iii)* Of the 11 *E. coli* insertions that did not occur in repeat regions, 7 can be explained by backward-slippage followed by re-transcription of the slipped bases. And for the remaining 4 in which the inserted nucleotides that do not match the bases preceding the insertion, there were no sequence characteristics signifying the source of the error.

Unlike backward slippage-events, the majority of which occurred in runs of A or T, forward slippage events, which result in transcription deletions, were not dependent on homopolymeric repeats and were significantly more likely to occur when the most recently transcribed two bases were CA, UA, or AU in *E. coli*. Additionally, our finding

that guanine is underrepresented before and after a transcription deletion aligns with a finding that guanine is enriched at the −2 and +1 sites in pause-prone sequences (146), implying that G-rich sequences stimulate pausing whereas G-poor sequences are slippery.

These new transcriptome-wide data support a revised model of transcription slippage in which increased RNA:DNA hybrid complementarity after slippage fosters the elongation complex to resume transcription at a new site, resulting in a transcription insertion or deletion. Previous models implied that the occurrence of transcription slippage was limited to homopolymeric runs; and we now conclude that it is the overall complementarity of the RNA:DNA hybrid after transcription slippage that contributes to the creation of indels.

When a ribonucleotide is misincorporated during transcription, the unpaired base causes the transcript to bend away from the template ("fraying") (50), which induces the elongation complex to pause and translocate backwards while extruding the unpaired base—a process termed "backtracking" (48, 62). If a transcription slippage event results in mispairing at the 3'-end of the RNA:DNA hybrid, it may resemble a misincorporation, causing the elongation complex to attempt to backtrack. Because the transcript and elongation complex reside in a new location after slippage, backtracking will be blocked because this process requires complementary base-pairing between the transcript and the DNA template (48, 62). If a portion of the nascent transcript slips forward and through the elongation complex before slippage stops, the resulting RNA:DNA hybrid can resemble a backtracked state, such that nucleolytic cleavage might still occur. However,

the orientation of nascent transcripts relative to the elongation complex cannot be inferred from our assays, so the effect of nucleolytic cleavage on slipped transcripts presently remains unknown.

Because the stability of the elongation complex is affected by the RNA:DNA hybrid, low complementarity after slippage may cause the slipped elongation complex to dissociate (48, 138, 147). However, if upstream elongation complexes collide with the slipped elongation complex before it dissociates due to poor base complementarity, it may be advanced forward to a region of high RNA:DNA hybrid complementarity so that transcription can resume. If the forward action of upstream elongation complexes is the primary mechanism of forward slippage, the distance that an elongation complex can be pushed before it dissociates may dictate the maximum length of transcription deletions.

The mechanisms that generate indels during replication are similar to those resulting in transcription slippage, but there are fundamental differences between the processes: First, the majority of indels in DNA occur in short repetitive regions (148), whereas those in RNA transcripts occur in more complex sequences. In DNA, indels are thought to be generated by slipped-strand misalignment (148), and our model of transcription indels involves a similar mechanisms but does not require the presence of direct repeats. Second, small indels in DNA can be generated through dNTP-stabilized misalignment (149), whereas a similar mechanism occurring during RNA transcription would produce base substitutions. The difference is due to the manner by which DNA

polymerase and RNA polymerase handle the conformational constraints of a displaced base (150, 151).

Overall, our model of transcription slippage (Figure 3.9 and Figure 3.10) involves two steps that lead to transcription insertions and deletions: First, the transcript RNA loses register with the template DNA, causing the elongation complex to slip along the template DNA. The amount of slippage is influenced by presence of upstream elongation complexes, which can block extensive backward-slippage and even propel the slipping elongation complex to a new location. Next, slippage events that result in high RNA:DNA hybrid complementarity, particularly at the 3' end, lead to re-initiation of transcription elongation to generate an insertion or deletion. Whereas our model is based on the sequence locations at which indels occur, additional experimental work is required to determine the accuracy of the proposed mechanism.

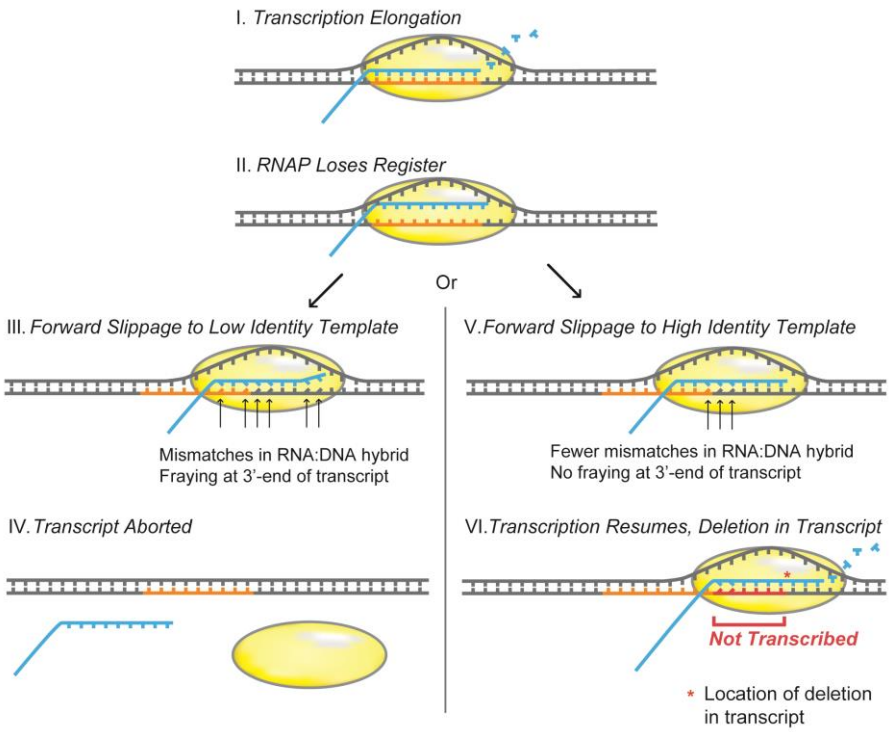

Figure 3.9 – Model of transcription slippage resulting in deletions

Based on locations and sequence contents of deletions genome-wide, the degree complementarity of RNA:DNA hybrid after a transcription slippage event (I and II) determines whether transcription is aborted, producing a truncated transcript (III and IV), or resumed, producing a transcript containing a deletion (V and VI). Steps in the model use the following notation: Template DNA is shown in black, transcript RNA and incoming ribonucleotides in blue, the original RNA:DNA hybrid location is orange, the non-transcribed (*i.e.*, deleted) region in red, mismatched bases as angled contacts between non-complementary nucleotides, and the RNAP transcription elongation complex is represented by a yellow bubble. In this model, normal transcription (I) becomes interrupted when the elongation complex and transcript lose register with the DNA template (II). Possible outcomes include, the elongation complex slipping forward to a region of low complementarity (III), and in this example depicted, the elongation complex slips forward five bases, landing on a template location where six of the nine bases in the RNA:DNA hybrid are not complementary. If transcription cannot resume due to the extent of mispairing in the RNA:DNA hybrid and/or fraying at the end of the transcript, the transcript is aborted (IV). Alternatively, if the elongation complex slips to template location with fewer mismatches (V), the 3'-end of the RNA bonds sufficiently to the DNA template, and transcription resumes (VI) after the skipped the region, generating a deletion.

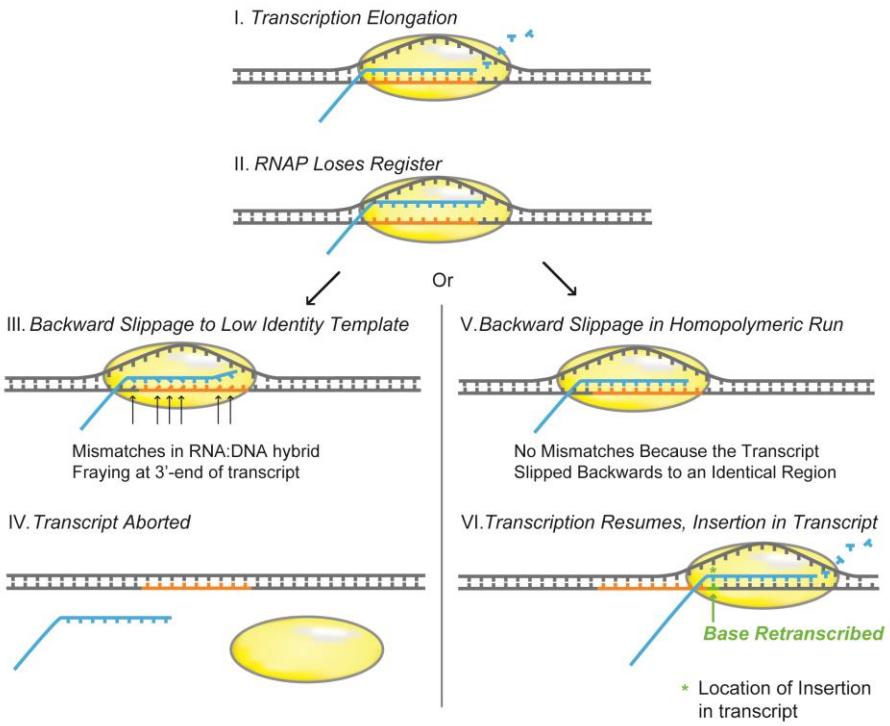

Figure 3.10 – Model of transcription slippage resulting in insertions

Based on locations and sequence contents of insertions genome-wide, the degree of complementarity of the RNA:DNA hybrid after a transcription slippage event (I and II) determines whether transcription is aborted, producing a truncated transcript (III and IV), or resumed, producing a transcript containing an insertion (V and VI). Steps in the model use the following notation: Template DNA is shown in black, transcript RNA and incoming ribonucleotides in blue, the original RNA:DNA hybrid location is orange, the re-transcribed (*i.e.*, insertion) region in green, mismatched bases as angled contacts between non-complementary nucleotides, and the RNAP transcription elongation complex is represented by a yellow bubble. In this model, normal transcription (I) becomes interrupted when the elongation complex and transcript lose register with the DNA template (II). Possible outcomes include, the elongation complex slipping backward to a region of low complementarity (III), and in this example depicted, the elongation complex slips backward one base, landing on a template location where six of the nine bases in the RNA:DNA hybrid are not complementary. If transcription cannot resume due to the extent of mispairing in the RNA:DNA hybrid and/or fraying at the end of the transcript, the transcript is aborted (IV). Alternatively, if the elongation complex slips to template location with fewer mismatches, in this case a homopolymeric run (V), the 3'-end of the RNA bonds sufficiently to the DNA template, and transcription resumes (VI) after the re-transcribed region, generating an insertion.

3.5 MATERIALS AND METHODS

3.5.1 Strain Information, sequencing procedures, and detection of indels

We assayed the transcriptomes of eight biological replicates of *Escherichia coli* MG1655 and two biological replicates of *Buchnera aphidicola* LSR1 using the CirSeq library preparation protocol (14). In this method, mRNA is sheared into 80–100 bp fragments, which are then circularized, primed using random hexamers, and reversed transcribed to generate cDNA that contains multiple linked repeats of the mRNA fragment. cDNAs containing these repeats were sequenced using Illumina MiSeq 300-nt read-lengths to capture at least three repeats within a sequencing read. Reads were processed by the CirSeq_v3 pipeline to generate a consensus sequence for each read (13) (http://andino.ucsf.edu/ CirSeq). All settings used in CirSeq_v3 were default with a quality score cutoff of 20. CirSeq_v3 uses Bowtie 2 (152) to align reads to a reference genome (NC_000913.3 for *E. coli* and NZ_ACFK01000001 for *Buchnera*). Additionally, we edited the 'run.sh' script to retain the intermediate output (9_alignment.sam and 10_alignment.sam) generated during the CirSeq_v3 pipeline, since they contain candidate insertions and deletions. Additional strain information and library preparation protocols have been described elsewhere (3). The data are publicly available from the NCBI SRA (SRP072992).

By generating a consensus sequence from the multiple repeats within a single read, sequencing errors, which appear as changes in only one of the repeats, are omitted. Insertion and deletion rates of Illumina sequencing are very low (153), and only those

insertions or deletions that occur at identical positions and are of equal size in fully aligned repeats were considered authentic. Because sequencing reads originate from the reverse transcription of circularized mRNA fragments primed with random hexamers, the actual orientation of sequences can only be determined after multiple rounds of sequence alignment. This process generates intermediate alignment files (9_alignment.sam and 10_alignment.sam) that contain many improperly mapped reads, and to detect insertions and deletions, we searched these files to identify reads that contained indels flanked on both sides by fully aligned sequences. One strategy for determining the correct orientation of a read in the CirSeq_v3 pipeline was to sequentially move each base from one end of the read to the other (13). By mapping each iteration to the genome, many reads that initially contained insertions or deletions eventually yielded an aligned sequence devoid of indels. To identify insertions and deletions, we retained those reads that contained the highest alignment score within each iteration of a read while also containing an insertion or deletion. Finally, only those insertions receiving quality scores ≥20 and only those deletions that were flanked on both sides by bases receiving quality scores ≥20 were considered. Additionally, we sequenced the genome of the parental strain of *E. coli* to confirm that no errors were attributable to genomic mutations. Statistical analyses were performed with Prism GraphPad and R.

3.5.2 Simulations

To determine if the observed deletions are biased toward specific sequences, we calculated their expected occurrence through simulations based on the frequencies of

gene transcripts in the transcriptome. The average read depth of each gene was tabulated, and genes were sampled at random, weighted by read depths. Because there was no observed bias in the locations of deletions within genes, simulated deletions could be allowed to randomly occur anywhere within the coding region of a transcript. The length of each deletion was drawn from the distribution of deletion lengths for each replicate without resampling. We performed 100 replicate simulations for each transcriptome examined. All simulations were subjected to the same adjustments and analyses (described below) as the observed deletions.

3.5.3 Ascertaining locations and contents of deletions

In many cases, it is possible to identify the precise location of a deletion by aligning reads to the reference sequence; however, many deletions occurred in regions of low sequence complexity or involved the deletion of a repeat in a repetitive sequence. Such cases can result in ambiguities in ascertaining which of the multiple, identical repeat-units was deleted, so these were resolved by positioning the ambiguous portion to the 3'-end of the deletion. Because this procedure may artificially increase base complementarity at the 3'-end of reconstructed RNA:DNA hybrids (see below), we controlled for any introduced biases by treating simulated deletions in the same manner.

3.5.4 Computing indel rates

Transcriptome-wide rates of insertions and deletions of *E. coli* and *Buchnera* were calculated for each replicate by dividing the total number of insertions or deletions in

protein-coding transcripts by the sequencing coverage of the corresponding regions, and averaging across replicates. To calculate the indel error rates at homopolymeric runs, we first identified all homopolymeric runs ≥4 nucleotides in length within protein-coding genes of *E. coli* MG1655 and *B. aphidicola* LSR1, and then determined the numbers of insertions and deletions originating in runs of each length category. When evaluating error rates in homopolymeric runs, or across any gene category, indel frequencies were normalized to the sequence coverage for each category. To determine the effect of transcript abundance on error rate, all genes were binned by their average coverage, and the errors and total coverage were tabulated for each bin. Coverage bins increased in 1x increments from 0–10-fold coverage; 10x increments from 10–100-fold coverage; 100x increments from 100–500-fold coverage; and subsequently in 500–1000-fold, 1000–2000-fold, and >2000-fold coverage bins. Transcriptomes were analyzed using custom python scripts, and all statistics were performed using Prism GraphPad and R.

3.5.5 Features of deleted regions

The nucleotide compositions of deleted nucleotides, and for the 15-bp regions preceding and succeeding each deletion, were calculated by direct count for each observed or simulated replicate, and then pooled across replicates. We inferred the complementarity of bases within the RNA:DNA hybrids after a slippage event by comparing the nine nucleotides directly preceding the start of each deletion to the nine nucleotides directly preceding the end of each deletion. The nine nucleotides preceding the start of each deletion represent the nucleotides transcribed before the slippage event

and constitute the RNA portion of the RNA:DNA hybrid, and the nine nucleotides preceding the end of each deletion represent the region in which slippage stopped and constitute the new portion of DNA in the RNA:DNA hybrid.

Chapter 4: A genome-wide assay specifies only GreA as a transcription fidelity factor in *Escherichia coli*[1]

4.1 ABSTRACT

Although mutations are the basis for adaptation and heritable genetic change, transient errors occur during transcription at rates that are orders of magnitude higher than the mutation rate. High rates of transcription errors can be detrimental by causing the production of erroneous proteins that need to be degraded. Two transcription fidelity factors, GreA and GreB, have previously been reported to stimulate the removal of errors that occur during transcription, and a third fidelity factor, DksA, is thought to decrease the error rate through an unknown mechanism. Because the majority of transcription-error assays of these fidelity factors were performed *in vitro* and on individual genes, we measured the *in vivo* transcriptome-wide error rates in all possible combinations of mutants of the three fidelity factors. This method expands measurements of these fidelity factors to the full spectrum of errors across the entire genome. Our assay shows that GreB and DksA have no significant effect on transcription error rates, and that GreA only influences the transcription error rate by reducing G→A errors.

---

[1] This chapter is reproduced (with minor modifications) from its initial publication:

Traverse CC and Ochman H (2018) A Genome-wide assay specifies only GreA as a transcription fidelity factor in *Escherichia coli*. *G3 (Bethesda)* 8:2257–2264. Ochman H supervised the project.

4.2 INTRODUCTION

All organisms are subject to non-heritable errors that are introduced into RNA during transcription. Although these errors are transient, they contribute considerable variation to the proteome and in the modification of proteins sequences; and in humans, these errors have been associated with aging and the development of cancer (1). In bacteria, transcription errors occur orders of magnitude more frequently than mutations in DNA and are prevalent across the entire transcriptome (2–5, 11). It has been estimated that about 1 in 10 proteins would be altered due to the high rate of transcription errors (Chapter 2 of this dissertation) (3). Although these transient errors have been hypothesized to have some benefit under stressful conditions (7–9, 77), most are probably deleterious and generate harmful or non-functional protein variants that need to be degraded.

In addition to variant proteins that originate from transcription errors, misincorporations can stall RNAP to interfere with DNA replication (36, 37, 134, 135). When an error occurs during transcription, the misincorporated base triggers the RNAP to halt transcription and translocate backwards along the DNA template while simultaneously extruding the error from the RNAP, a process called “backtracking” (48, 154). If this backtracked RNAP is not resolved, RNAPs can accumulate upstream, posing a barrier to DNA replication enzymes and generating double-strand breaks (35–37, 134, 135). To mitigate the effects of transcription errors, bacteria have evolved quality-control strategies that serve to restart backtracked RNAP: The RNAP can either undergo intrinsic

cleavage, whereby the RNAP itself catalyzes the removal of the misincorporated base (46, 63, 155, 156), or the error can be removed by Gre-mediated cleavage, in which secondary proteins bind to the RNAP and induce transcript cleavage (157–159).

Two Gre proteins, GreA and GreB, restart paused RNAPs by resolving backtracked RNAP and, as a result, resolve errors that prompted the RNAP to pause. These proteins are considered to be transcription fidelity factors (or anti-backtracking factors) since they have been shown to remove misincorporations in *in vitro* transcription assays and *in vivo* reporter gene assays (67, 69, 85, 159). Recently, a sequencing-based study recognized a role for GreAB in reducing G→A errors (90); however, that methodology is prone to sequencing artefacts, even after strict quality control. Additionally, that sequencing study measured the nascent transcripts that reside within paused RNAP, some of which may not have undergone intrinsic or Gre-mediated cleavage. Consequently, the effects of GreAB on the rates and profiles of errors that are incorporated into the transcriptome remain unexplored.

Recently, DksA, which competes for the same binding site as GreA and GreB on the RNAP, has been identified as a third transcription fidelity factor based on *in vivo* and *in vitro* assays (75, 76). DksA, which is structurally similar to GreA and GreB, does not induce transcript cleavage but instead reduces the occurrence of transcription errors through an unidentified mechanism (10). Moreover, the error rate and the types of errors prevented by DksA remain unknown. In this study, we employ a technique that eliminates sequencing artefacts (13, 14), and has allowed us to advance the measurement

of transcription error rates to all types of substitutions, including base substitutions and indels, across the entire transcriptome. Our assay found no effect of GreB and DksA on the transcription error rate, and that GreA reduces only the rate of G→A errors, as previously reported (90). These results suggest that intrinsic cleavage, although slow, may have a larger role in resolving misincorporated bases than previously expected.

4.3 RESULTS

4.3.1 GreA appears to be the sole transcription fidelity factor

To determine the effects of GreA, GreB, and DksA on transcriptional fidelity, we used a transcriptome-wide sequencing approach that discriminates sequencing artefacts from actual errors that arose during transcription by circularizing mRNAs, reverse-transcribing the circularized fragments, and sequencing cDNAs that contain multiple linked repeats of the original mRNA fragment (13, 14). A consensus sequence is then calculated from the repeats to recognize errors arising during library preparation and sequencing (which only occur once per repeat) from errors that were present in the original mRNA fragment (which appear in every repeat). Applying this method to measure the transcription error rate in mutant strains lacking one, or any of the possible combinations, of these genes (including the triple mutant), yielded no mutant strains that differed significantly from one another or from the wildtype (Figure 4.1A; unpaired Student's *t*-tests, n = 2, $p > .2$). However, there was a tendency for mutants lacking the *greA* gene (*i.e.*, Δ*greA*, Δ*greAgreB*, Δ*greAdksA*, Δ*greAgreBdksA*; red-shaded points in Figure 4.1) to have slightly higher error rates than strains that possessed an intact *greA*

gene, even in combination with a deletion in one or both of the other fidelity factors (*i.e.*, MG1655, Δ*greB*, Δ*dksA*, Δ*greBdksA*; blue-shaded points in Figure 4.1). By grouping strains based on their possession or lack of *greA*, the transcription base substitution rate was significantly higher in Δ*greA* strains (Figure 4.1B, Mann-Whitney U-test, n = 8, p = .007), indicating that GreB and DksA do not contribute to overall transcriptional fidelity under the conditions tested.

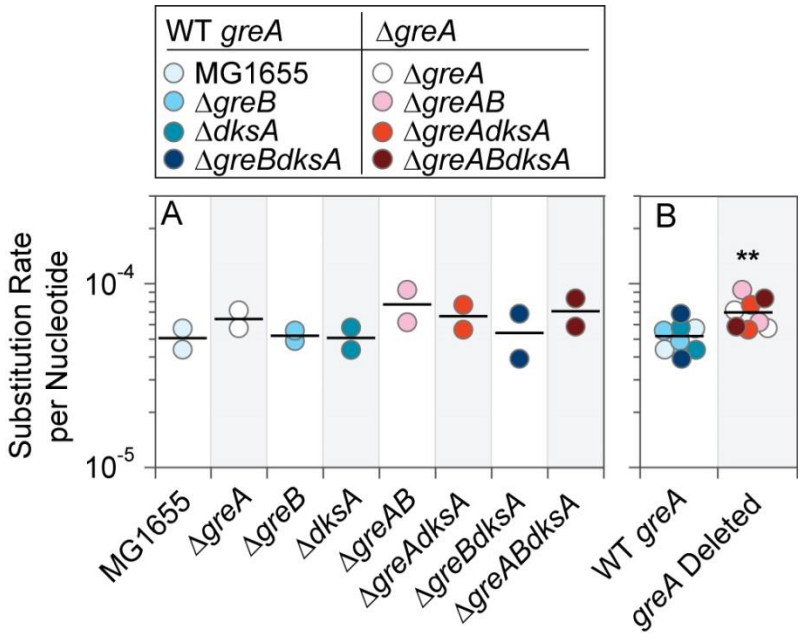

Figure 4.1 – Transcription error rates in *E. coli* strains lacking one or multiple fidelity factors.

A. Rates of transcription base substitutions in wildtype *E. coli* strain MG1655 and in isogenic strains harboring deletions of all possible combinations of three fidelity factors, *greA*, *greB*, and *dksA*. There are no significant differences of the transcription substitutions rates between wild-type *E. coli* MG1655 and any the fidelity factor mutants (unpaired Student's *t*-tests, n = 2, p > .2). **B.** Rates of transcription substitutions of all strains with an intact *greA* gene (blue-shaded points) and all mutants lacking the *greA* gene (red-shaded points). The overall error rate in Δ*greA* strains is significantly higher than in strains with wild-type *greA* (Mann-Whitney U-test, ** p < .01). The same y-axis is used as in **A**.

4.3.2 GreA only corrects G→A substitutions

We next sought to determine if specific base substitutions were differentially affected by each of the transcription fidelity factors. In those mutant strains that harbored an intact *greA* (Δ*greB,* Δ*dksA,* Δ*greBdksA*), there were no significant effects on the error rates of individual substitutions (Figure 4.2); however, G→A substitutions were significantly higher in all Δ*greA* strains (Figure 4.2). This trend remains when all statistical tests were performed on strains grouped according to whether or not they possessed an intact *greA* gene, an intact *greB* genes, or an intact *dksA* gene (Figure 4.3, $p = .0001$, Mann-Whitney U-test corrected by Benjamini-Hochberg procedure).

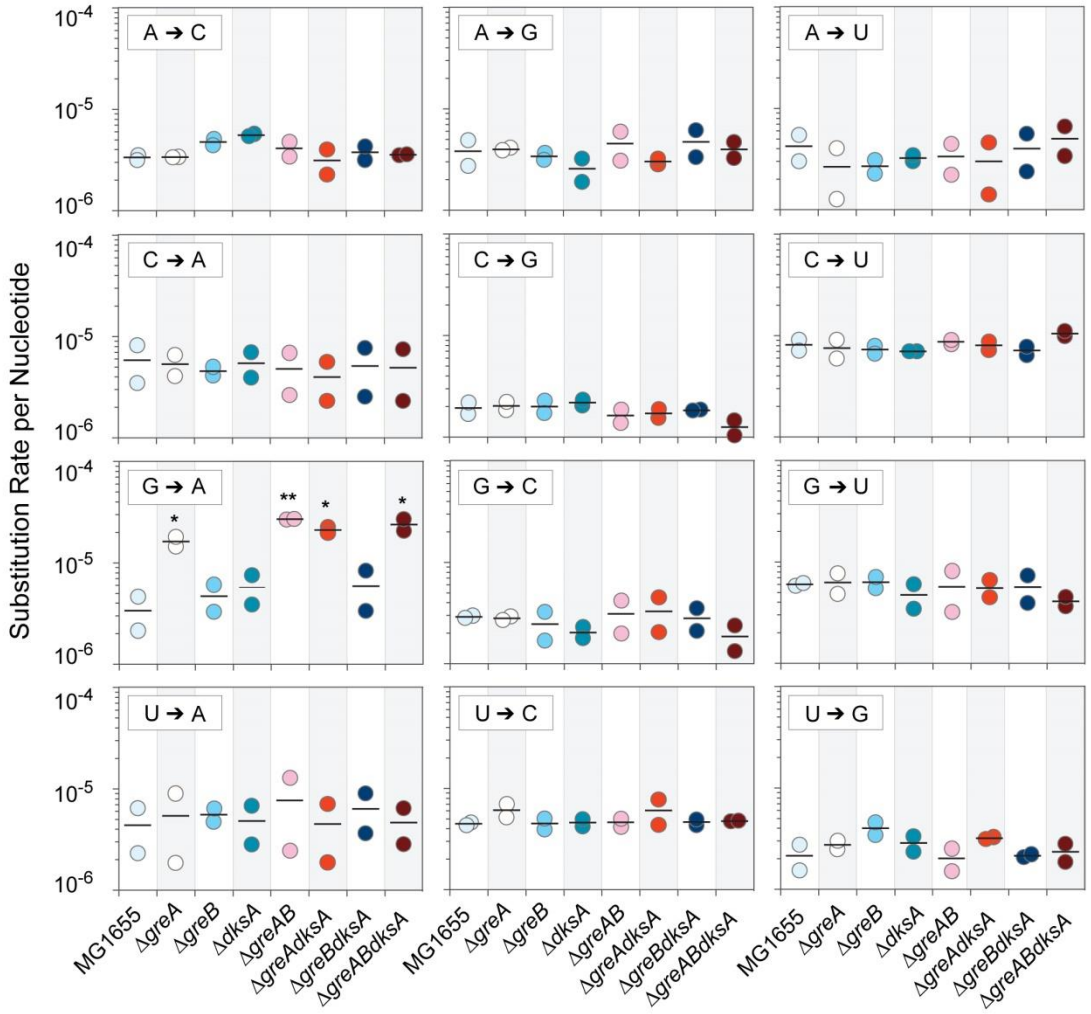

Figure 4.2 – Transcription error rate for each type of base substitution in wild-type *E. coli* MG1655 and each fidelity factor mutant

Each of the mutant strains with *greA* deleted have a significantly higher G→A substitution rate than wild-type *E. coli* MG1655 (unpaired Student's t-tests: Δ*greA*, p = .027; Δ*greAgreB*, p = .003; Δ*greAdksA*, p = .011; Δ*greAgreBdksA*, p = .027). No other comparisons were statistically significant. All tests were subject to correction for multiple tests by the Benjamini-Hochberg procedure. * $p < .05$; ** $p < .01$.

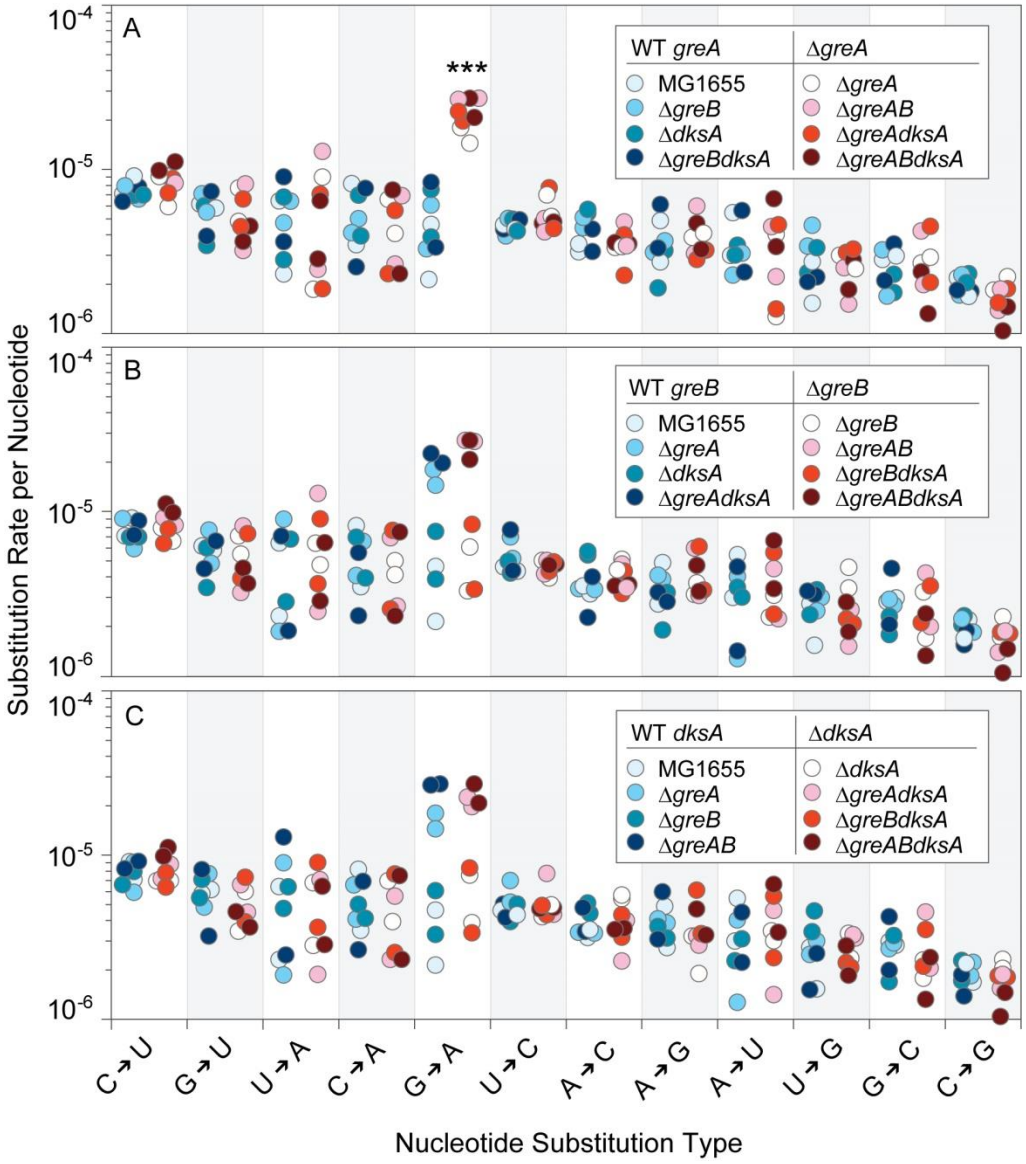

Figure 4.3 – Transcription error rates for each substitution type grouped each fidelity factor

The transcription substitution rates were calculated when the replicates were grouped by **A** wild-type *greA* and Δ*greA*, **B**, wild type *greB* and Δ*greB*, and **C**, wild-type *dksA* and Δ*dksA*. The transcription error rate of G→A substitutions in Δ*greA* replicates was significantly higher than wild-type *E. coli* MG1655 (Mann-Whitney test, n = 8, p = .0001). No other comparisons were statistically significant for any of other mutant groupings. All tests were subject to correction for multiple tests by the Benjamini-Hochberg procedure. *** $p < .001$.

4.3.3 Cytosine is overrepresented prior to G→A errors

During transcription, the nine most recently transcribed bases remain hybridized to the template DNA within the RNAP (known as the RNA:DNA hybrid), and previous work has suggested that these bases may influence the error rate (90). To determine if the most recently transcribed RNA influences the error rate, we analyzed the occurrence of each of the four nucleotides in bioinformatically reconstructed RNA:DNA hybrids (Chapter 3 of this dissertation) (160) immediately preceding each of the observed errors. We found that cytosine was significantly overrepresented in the position immediately preceding a transcription error in Δ*greA* mutant strains (Figure 4.4). We examined this in further detail by analyzing how each of the four nucleotides influenced the error rate for each substitution type (Figure 4.5). We found that G→A substitutions were significantly more likely to occur if any nucleotide but A preceded the substitution in the Δ*greA* mutant, with the strongest effect produced by C. No other nucleotide preceding any of the other type of base substitution significantly increased or decreased the error rate.

Our analysis focused on errors that resulted in base substitutions, but transcription errors can also produce insertions or deletions. None of the mutant strains, or groupings of strains, displayed a significant effect on transcription indel rates, nor did they cause differences in errors according to the strand or genomic location of transcription, or the level of gene expression.

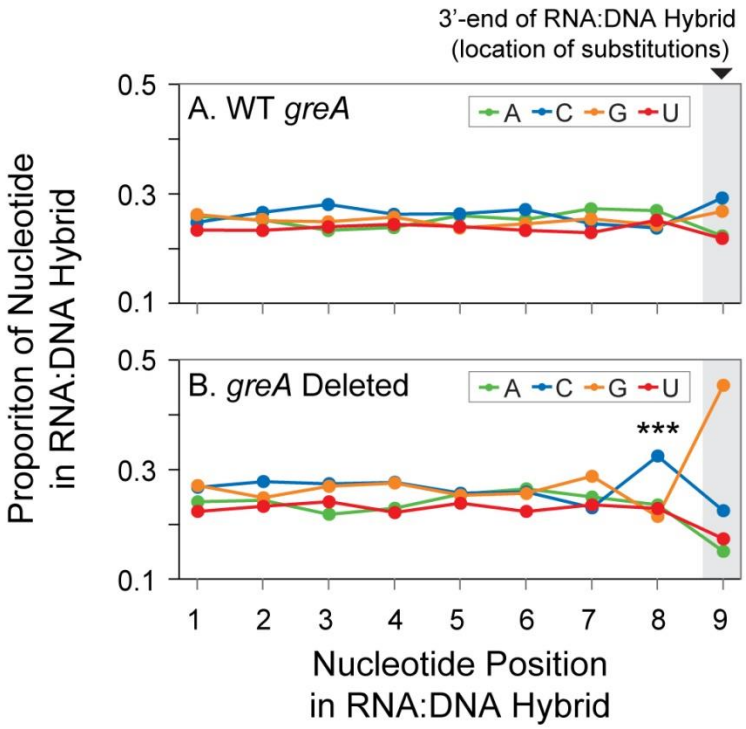

Figure 4.4 – Nucleotide composition in the RNA:DNA hybrid at positions preceding a transcription error

The proportion of each nucleotide at each position within the RNA:DNA hybrid was calculated for all strains with an intact wild-type *greA* gene and in which the *greA* gene was deleted. The shaded gray area marks the 3'-end of the RNA:DNA hybrid at the site where the transcription error occurred. In strains lacking *greA*, the occurrence of C was significantly higher in the position immediately before a transcription error (Fisher's exact test, *** $p < .0001$), and no other positions in the RNA:DNA hybrid exhibit a significant difference in nucleotide composition between strains. The results for each position were normalized by the base composition of the sequenced transcriptome. All tests were subject to correction for multiple tests by the Benjamini-Hochberg procedure.

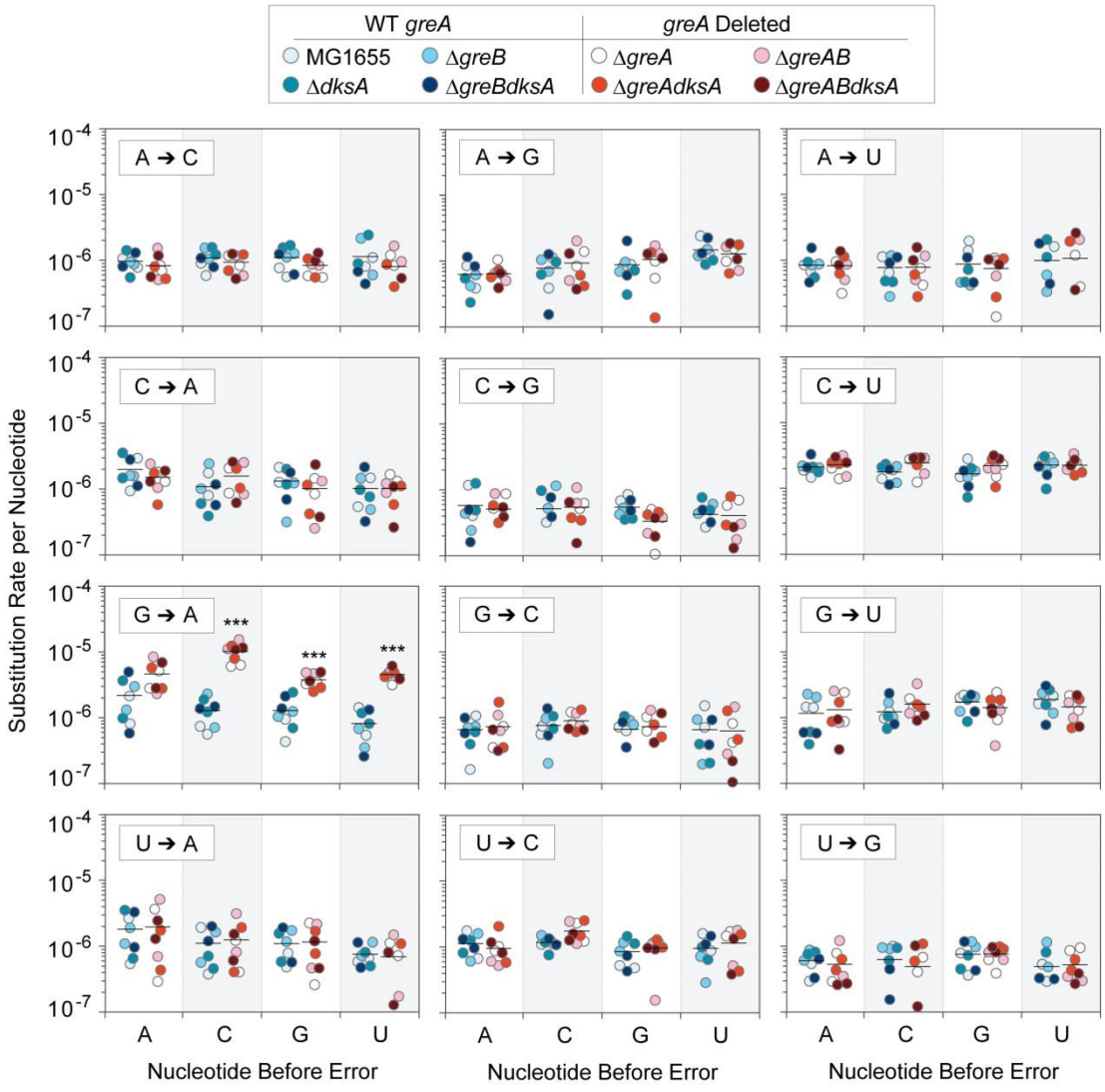

Figure 4.5 – Effect of preceding nucleotide on error rates of each substitution type

The transcription error rate was calculated for all replicates with wild-type *greA* and all Δ*greA* replicates when each nucleotide occurs before each substitution type. The G→A substitution rate is significantly higher for Δ*greA* replicates when preceded by C, G, and T (Mann-Whitney tests, n = 8, $p < .0002$ when preceded by C, G, and T). All tests were subject to correction for multiple tests by the Benjamini-Hochberg procedure. *** $p < .001$.

4.4 DISCUSSION

Three transcription fidelity factors—GreA, GreB, and DksA—have been described in *E. coli* (67, 69, 159), and by applying a transcriptome-wide approach that registers all errors suppressed by these factors (13, 14), we conclude that, under the conditions tested, GreB and DksA do not significantly influence transcriptional fidelity and that GreA reduces only the G→A error rate. Indeed, a recent study used circle sequencing to demonstrate that the *Saccharomyces cerevisiae* TFIIS gene, the eukaryotic homolog of GreA, reduces the G→A error rate more than all other errors (161), indicating that the preponderance of G→A errors in mutants lacking fidelity factors may be universal. Our finding that the other recognized fidelity factors are of little consequence in correcting transcription errors counters previous views on GreB-mediated cleavage. Prior work has suggested that GreB increases transcription fidelity *in vitro* (162); however, further support for the action of GreB on transcription fidelity has been extrapolated either from its ability to cleave backtracked transcripts (10, 66, 69, 163) or from studies that test Δ*greAB* mutants and cannot disentangle the individual contributions of the two proteins (2, 90).

Recently, information on RNAP pausing from an alternate transcriptome-wide approach, termed NET-seq (164, 165), was used to examine the effect of Δ*greAB* mutants on rates of transcript misincorporation (90). NET-seq captures transcript sequences that reside within the RNAP (i.e., before most error correction can occur) and yields error rates that are orders of magnitudes higher than we obtained when surveying transcripts

that have been released from the RNAP. The difference between these rates is that the estimates obtained through NET-seq can include errors that have not yet undergone intrinsic cleavage as well as those in transcripts that are eventually aborted and are not part of the mature transcriptome. In line with our results, only the G→A error rate substantially increased in the Δ*greAB* mutant when assayed by NET-seq, although it was not determined if the effect was attributable solely to GreA (90).

We also found evidence of biases in bases preceding certain errors. NET-seq found the C was more likely to be transcribed prior to a G→A error and we found similar results with CirSeq: C had the largest effect on the G→A error rate, but G and T were also elevated prior to G→A errors. The mechanism underlying the increase of C nucleotides immediately preceding a G→A error is unclear from our results. For example, using our methodology, it is not possible to determine if all errors increase subsequent to transcription of cytosine but intrinsic cleavage is able to correct all errors except for G→A, or if only G→A errors are increased following cytosine. It is possible that the 3'-nt structure of A (misincorporated opposite of C) influences either the intrinsic cleavage of the misincorporated nucleotide or the ability of the RNAP to detect the misincorporation event. However, previous *in vitro* work does not indicate that G→A is harder to resolve through intrinsic cleavage than N→A errors (63), but these measurements did not take into account all possible preceding nucleotides.

If NET-seq only registered transcripts prior to error correction, it would yield the same error rates for wild-type and Δ*greAB* mutants, due to the fact that Gre acts on

transcripts after misincorporation. That the G→A error rate increases in Δ*greAB* mutants relative to wild-type indicates that NET-seq interrogates not only those transcripts that never experienced an error and those that have not undergone intrinsic or Gre-mediated cleavage, but also those that have already undergone intrinsic or Gre-mediated cleavage (90, 166). A previous study concerning *Thermus aquaticus* RNAP has shown that intrinsic cleavage mechanisms remove misincorporations involving adenine at much faster rates than other misincorporations (63), and consequently, the actual input of G→A errors is likely higher than the 10-fold increase reported for the Δ*greAB* mutant assayed by NET-seq. Although G→A errors should be removed by intrinsic cleavage at a faster rate than other errors (63), it appears that the input of these errors is so high that it requires the additional action of Gre-mediated cleavage. It is important to note that intrinsic cleavage has been measured *in vitro* as being very slow, and consequently, intrinsic cleavage was not thought to significantly contribute to transcription fidelity. However, the low error rates that we obtained suggest that intrinsic cleavage may operate at a faster rate *in vivo* or that there is possibly an as-yet unidentified cleavage factor.

The NET-seq findings support our results, but they only assayed a double mutant and did not separate the individual effects of GreA and GreB. We find that GreB does not act on any class of transcription errors, which is inconsistent with prior findings (162) and views (10, 66, 69, 163) on GreB-mediated cleavage. However, a recent study that used an *in vivo* reporter system to specifically probe G→A errors reported that GreA, and not GreB, affected the G→A error rate (167), but that overexpressing GreB in the Δ*greA* mutant could mitigate G→A errors. Because GreB operates on transcription errors only

under atypical conditions (*i.e.*, at very high concentrations in strains lacking *greA*) suggests that GreA is the major fidelity factor and implies that GreB has a separate function (67, 69, 167).

The difference between the results obtained for GreA and GreB can be traced to their roles in inducing cleavage in RNAPs that have backtracked by different lengths: GreA preferentially associates with backtracks of only 2 or 3 bases, whereas GreB associates with backtracks up to 18 bases in length (65–69). And because most misincorporations that occur during transcription induce short backtracking events (46, 63, 168), GreA will be the dominant, if not sole, fidelity factor detected by *in vivo* systems. GreA and GreB were originally classified as transcription fidelity factors due to their ability to induce nucleolytic cleavage of misincorporated transcripts; however, they also serve as anti-backtracking factors that prevent DNAP-RNAP collisions (36, 37, 134). Therefore, GreA may not increase fidelity *per se* but instead may restart backtracked RNAP, such that increased fidelity is a consequence of restarting transcription.

The third fidelity factor tested was DksA, which is known to have a role in transcription initiation (70, 72, 169, 170), elongation (35), and genome stability (36, 37, 134). DksA and Gre have similar structures and RNAP binding locations, but unlike Gre, DksA does not induce nucleolytic cleavage (171). Whereas a study showed that DksA reduces transcript read-through by inhibiting misincorporations *in vitro* and *in vivo* (75), this error avoidance mechanism is not observed in our assay. Additionally, a Δ*dksA* mutant increases the readout of transcription errors in a reporter assay (76); however,

transcription errors were not measured directly such that error rates could not be derived. The discrepancies between our transcriptome-wide analyses and these assay systems suggest a subtle role for this protein that possibly occurs below our limit of detection or under conditions not tested, such as during the stringent response (where ppGpp could act synergistically with DksA) (75, 171) or the general stress response (35, 36). Under such conditions, transcription and translation can become uncoupled, and when RNAP and the ribosome do not physically interact, RNAP is prone to pausing (35). Although misincorporations induce RNAP pausing (90, 172) and this pausing is known to be mitigated by DksA (35), the degree to which this protein helps prevent errors across the transcriptome is not yet evident.

Therefore, of the three previously identified fidelity factors, only GreA appears to act as a fidelity factor. Because we only tested the roles of GreA, GreB, and DksA under a single condition, it is important to note that they could possibly affect transcription fidelity under other assay conditions (e.g., stationary phase, stringent response, general stress response, etc.). Furthermore, the ~100-fold difference between our reported G→A error rates in Δ*greA* mutants and those reported in Bubunenko *et al*. (2017) may stem from the different assay conditions: if the reporter-based assay induces stressful conditions, then the fidelity factors may become more important for error correction than in the conditions used in our study. Alternatively, this difference may stem from error rates that occur below our limit of detection. Although GreB and DksA may serve roles outside of error correction, our findings indicate that neither GreB nor DksA significantly influences transcription fidelity, as was found previously for GreB (167). Additionally,

intrinsic cleavage is considered a slow and inefficient mechanism of transcription error correction; however, we suggest that it may emend the majority of transcription misincorporations with additional action of GreA to remove G→A errors.

4.5 MATERIALS AND METHODS

4.5.1 Bacterial strains and growth conditions

All strains used in this study were derivatives of *Escherichia coli* MG1655. Mutant strains harboring deletions of *greA, greB* or *dksA* were supplied by M. Cashel (NIH), and new strain constructs harboring deletions in one, two, or three of these genes were generated with P1*vir*, as described previously (174). Bacteria were grown in LB to facilitate growth, avoid auxotrophies of the mutant strains, and because it has been shown that there are no differences in the transcription error rate when compared to growth in chemically defined minimal media (Chapter 2 of this dissertation) (3). Cultures and plates were supplemented with antibiotics as appropriate: chloramphenicol (Cm: 20 µg/ml), kanamycin (Kan: 40 µg/ml), and tetracycline (Tet: 20 µg/ml).

4.5.2 RNA extractions

For RNA extractions, newly transduced strains (to avoid the accumulation of suppressor mutations) were grown without antibiotics, and RNA was extracted during log-phase growth. RNA was isolated using the RNA*snap* protocol for gram-negative bacteria, as previously described (see methods in Chapter 2 of this dissertation) (3, 131). Ribosomal RNAs were removed from the total RNA preparations using the

MICROB*Express* kit (Life Technologies), according to manufacturer's instructions. Each sample represents an independent biological replicate that originated from independent cultures.

4.5.3 Library preparation and sequencing

The CirSeq method for preparing and sequencing RNA libraries was performed as described in Acevedo *et al*. (2014), with minor modifications (see methods in Chapter 2 of this dissertation) (3). Purified mRNA was mechanically sheared to 80–100 nt fragments, which were then fractionated and extracted by urea-PAGE. Isolated mRNAs were circularized, primed with random hexamers, and reverse transcribed, resulting in linked repeats of each original mRNA fragment. Resulting cDNAs were sheared into fragments 300–450 bp in length, and libraries prepared using the NEBNext Ultra RNA Library Prep Kit for Illumina sequencing (NEB). Samples were barcoded and sequenced on a MiSeq v3 platform generating 300-bp reads.

4.5.4 Data analysis

Sequences were processed using the CirSeq_v3 pipeline (13) to generate the consensus among the cDNA repeats within a sequencing read using default settings and a quality score cutoff of 20. Subsequent analyses were performed with the same custom python scripts previously described for the analysis of base substitutions (Chapter 2 of this dissertation) (3) and transcription indels (Chapter 3 of this dissertation) (160). The overall error rate was calculated by dividing the total number of transcription errors by

the total number of bases sequenced in the transcriptome. For individual error rates, the total number of errors for each error type was divided by the total number of bases sequenced in the transcriptome, such that the sum of all individual error rates is equal to the overall error rate. Additionally, the individual error rates were normalized by base composition, as previously described (Chapter 2 of this dissertation) (3). The error rates associated with nucleotides preceding a particular focal error were normalized by the nucleotide composition of positions −1 to −7 relative to each of the four bases. This was accomplished by randomly sampling the sequenced transcriptome one million times each for A, C, G, and T as the focal nucleotide and calculating the base composition for the eight bases preceding each sampled focal nucleotide. All statistics were performed in Prism Graphpad or R.

Chapter 5: Conclusions and future directions

The work presented in this dissertation represents the first comprehensive study of genome-wide transcriptional errors in bacteria. Although the rates of transcription errors have been measured for decades, the precision, accuracy, and scale of these measurements have been limited by the technologies available. The advent of high throughput sequencing provided the opportunity to study these errors across the entire transcriptome, except the error rate of RNA sequencing is too high to separate the sequencing errors from transcription errors. As discussed throughout the dissertation, there are studies that estimated the transcription error rate using sequencing data, but they are either limited to one location (2), are contaminated by sequencing artefacts (90), or miss the majority of errors because they are inefficient (89). The use of CirSeq in this dissertation improves on prior measures by overcoming each of these problems.

The capability to measure transcription errors in every transcript in the transcriptome allows us to test if specific sequence motifs, genomic loci, or locations in the transcript affect the error rate in any way. Under our conditions, we did not detect any transcription error rate differences in these different contexts. However, future experiments that test different stressful conditions or mutants may find biases that are currently masked in wild-type *E. coli* MG1655. Additionally, it is possible that slight differences in the transcription error rate do exist across the transcriptome, but we were either limited by our sample size or the sensitivity of CirSeq. Future improvements to both of these areas may yield new and interesting results.

The original CirSeq pipeline was not capable of detecting transcription insertions and deletions because the focus of the original study was to detect the substitution rate in RNA viruses (13). By altering the data pipeline, insertions and deletions are readily detectable in CirSeq data. Prior to this work, transcription slippage had only been studied in the context of homopolymeric runs (12). This genome-wide approach now reveals the propensity of RNAP to slip in more complex sequences than just homopolymeric runs. The finding that RNAP resumes transcription at sequences that most closely resemble the sequence it slipped from allowed us to develop a new and updated mechanistic model for transcription slippage. Future experiments could test this model *in vitro* using transcription assays or *in vivo* using mutagenesis, particularly in regions of the RNAP that interact with the RNA:DNA hybrid (12).

By applying CirSeq to *greA, greB,* and *dksA* mutants, we were able to directly test the contribution of these proteins to transcriptional fidelity. Our finding that only GreA affected the G→A error rate is perplexing, especially given the long history of the study of these proteins. However, the majority of research on these proteins was performed *in vitro* using concentrations of GreA or GreB that are not normally found *in vivo*. In this context, both GreA and GreB will restart misincorporated RNAP under these artificially high concentrations. Additionally, the majority of the work has been performed on GreA or GreA/GreB double mutants, and attributing GreB to the role of improving transcriptional fidelity may have been premature.

The finding that GreA only affects the G→A error rate may explain why the endosymbionts did not have elevated transcription error rates relative to *E. coli*. *Buchnera*

is only missing *greB* (Figure 2.1), which we found does not affect the transcription error rate *in vivo*, so it is not surprising that the error rate is not elevated. Despite *Carsonella* missing *greA*, *greB*, and d*ksA*, the only error that we would expect to increase would be G→A errors. Interestingly, G→A is the highest substitution rate in *Carsonella*, indicating that the lack of GreA may be responsible for this observation (Figure 2.6). However, there is only one sequenced replicate for *Carsonella*, therefore this result remains inconclusive.

The contribution of DksA to transcriptional fidelity is still unclear. Two separate studies show that DksA reduces transcription errors *in vitro* and in *in vivo*, but these experiments do not directly measure transcription errors (75, 76). Therefore, future experiments are needed to disentangle this discrepancy, perhaps by performing CirSeq on a double knockout of DksA and ppGpp synthesis, as was done *in vitro* (75). Additionally, DksA has been found to inhibit transcription pausing when transcription and translation are de-coupled, therefore the role of DksA in reducing transcription errors may become apparent under these conditions (35).

Finally, although this dissertation focused on protein-coding genes, non-protein coding genes can be measured too. When the transcription error rate of RNA genes (including tRNA and rRNA) are measured, they appear to have higher substitution rates than protein-coding genes (Figure 5.1). Initially, we thought this was an artefact of the method or an issue arising from multi-copy gene expression. However, James *et al*. (2016) reported a similar finding using a different library preparation method. Because RNA genes are not translated, there will not be transcription-translation coupling in these

transcripts. Transcription-translation coupling has been shown to reduce transcriptional pausing (34, 35) and base misincorporations are a major source of transcriptional pausing (90). The antitermination activity of the RNAP when transcribing RNA genes may allow for the RNAP to transcribe these genes in an error-prone manner to overcome the pausing that would occur without the physical interaction of a ribosome. To test if the lack of transcription-translation coupling is responsible for the increased error rate in RNA genes, future work could treat cells with serine hydroxamate (SHX) and measure the transcription error rate. Because SHX de-couples the ribosome from the RNAP, all new transcripts after SHX treatment would be synthesized without transcription-translation coupling. If the transcription error rate increases after SHX treatment, then transcription-translation coupling may be shown to increase the fidelity of transcription.

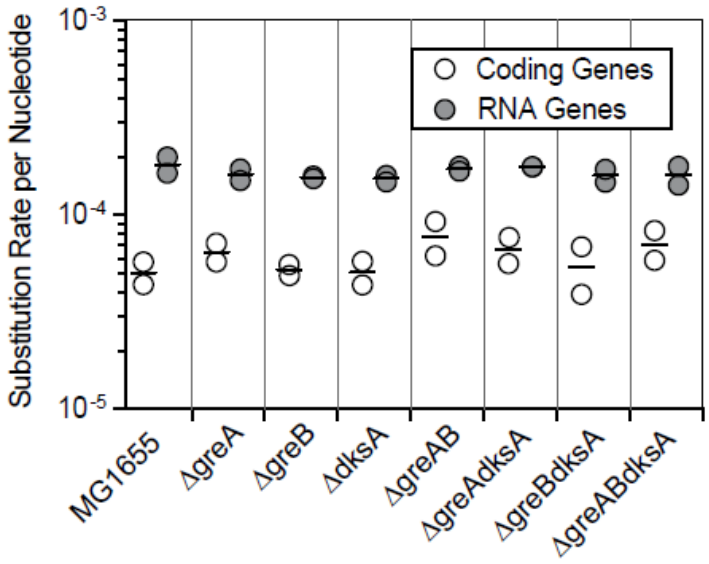

Figure 5.1 – Transcription substitution rate in protein coding genes and RNA genes

The substitution rate for protein coding genes (white circles) is compared to RNA genes (gray circles) in WT *E. coli* MG1655 and fidelity factor mutants. The RNA gene error rate is higher than protein coding genes in every sample.

References

1. Brégeon D, Doetsch PW (2011) Transcriptional mutagenesis: Causes and involvement in tumour development. *Nat Rev Cancer* 11:218–227.

2. Imashimizu M, Oshima T, Lubkowska L, Kashlev M (2013) Direct assessment of transcription fidelity by high-resolution RNA sequencing. *Nucleic Acids Res* 41:9090–9104.

3. Traverse CC, Ochman H (2016) Conserved rates and patterns of transcription errors across bacterial growth states and lifestyles. *Proc Natl Acad Sci USA* 113:3311–3316.

4. Rosenberger RF, Foskett G (1981) An estimate of the frequency of *in vivo* transcriptional errors at a nonsense codon in *Escherichia coli*. *Mol Gen Genet* 183:561–563.

5. Rosenberger RF, Hilton J (1983) The frequency of transcriptional and translational errors at nonsense codons in the *lacZ* gene of *Escherichia coli*. *Mol Gen Genet* 191:207–212.

6. Blank A, Gallant JA, Burgess RR, Loeb LA (1986) An RNA polymerase mutant with reduced accuracy of chain elongation. *Biochemistry* 25:5920–5928.

7. D'Ari R, Casadesús J (1998) Underground metabolism. *BioEssays* 20:181–186.

8. Meyerovich M, Mamou G, Ben-Yehuda S (2010) Visualizing high error levels during gene expression in living bacterial cells. *Proc Natl Acad Sci USA* 107:11543–11548.

9. Gordon AJE, *et al*. (2009) Transcriptional infidelity promotes heritable phenotypic change in a bistable gene network. *PLoS Biol* 7:e44.

10. Zenkin N, Yuzenkova Y (2015) New insights into the functions of transcription factors that bind the RNA polymerase secondary channel. *Biomolecules* 5:1195–1209.

11. Springgate CF, Loeb LA (1975) On the fidelity of transcription by *Escherichia coli* ribonucleic acid polymerase. *J Mol Biol* 97:577–591.

12. Zhou YN, *et al*. (2013) Isolation and characterization of RNA polymerase *rpoB* mutations that alter transcription slippage during elongation in *Escherichia coli*. *J Biol Chem* 288:2700–2710.

13. Acevedo A, Brodsky L, Andino R (2013) Mutational and fitness landscapes of an RNA virus revealed through population sequencing. *Nature* 505:686–690.

14. Acevedo A, Andino R (2014) Library preparation for highly accurate population sequencing of RNA viruses. *Nat Protoc* 9:1760–1769.

15. Vassylyev DG, *et al*. (2002) Crystal structure of a bacterial RNA polymerase holoenzyme at 2.6 Å resolution. *Nature* 417:712–719.

16. Lee DJ, Minchin SD, Busby SJ (2012) Activating transcription in bacteria. *Annu Rev Microbiol* 66:125–152.

17. Murakami KS (2013) X-ray crystal structure of *Escherichia coli* RNA polymerase σ^{70} holoenzyme. *J Biol Chem* 288:9126–9134.

18. Burgess RR (1969) Separation and characterization of the subunits of ribonucleic acid polymerase. *J Biol Chem* 244:6168–6176.

19. Bae B, Feklistov A, Lass-Napiorkowska A, Landick R, Darst SA (2015) Structure of a bacterial RNA polymerase holoenzyme open promoter complex. *Elife* 4:e08504.

20. Belogurov GA, Artsimovitch I (2015) Regulation of transcript elongation. *Annu Rev Microbiol* 69:49–69.

21. Nickels BE, Hochschild A (2004) Regulation of RNA polymerase through the secondary channel. *Cell* 118:281–284.

22. Kannan N, Chander P, Ghosh P, Vishveshwara S, Chatterji D (2001) Stabilizing interactions in the dimer interface of α-subunit in *Escherichia coli* RNA polymerase: A graph spectral and point mutation study. *Protein Sci* 10:46–54.

23. Murakami K, Kimura M, Owens JT, Meares CF, Ishihama A (1997) The two α subunits of *Escherichia coli* RNA polymerase are asymmetrically arranged and contact different halves of the DNA upstream element. *Proc Natl Acad Sci USA* 94:1709–1714.

24. Kohler R, Mooney RA, Mills DJ, Landick R, Cramer P (2017) Architecture of a transcribing-translating expressome. *Science* 356:194–197.

25. Vrentas CE, Gaal T, Ross W, Ebright RH, Gourse RL (2005) Response of RNA polymerase to ppGpp: Requirement for the ω subunit and relief of this requirement by DksA. *Genes Dev* 19:2378–2387.

26. Weiss A, *et al*. (2017) The ω subunit governs RNA polymerase stability and

transcriptional specificity in *Staphylococcus aureus*. *J Bacteriol* 199:e00459-16.

27. Ross W, Vrentas CE, Sanchez-Vazquez P, Gaal T, Gourse RL (2013) The magic spot: A ppGpp binding site on *E. coli* RNA polymerase responsible for regulation of transcription initiation. *Mol Cell* 50:420–429.

28. Feklístov A, Sharon BD, Darst SA, Gross CA (2014) Bacterial sigma factors: A historical, structural, and genomic perspective. *Annu Rev Microbiol* 68:357–376.

29. Paget MS (2015) Bacterial sigma factors and anti-sigma factors: Structure, function and distribution. *Biomolecules* 5:1245–1265.

30. Mooney RA, *et al*. (2009) Regulator trafficking on bacterial transcription units *in vivo*. *Mol Cell* 33:97–108.

31. Bakshi S, Choi H, Weisshaar JC (2015) The spatial biology of transcription and translation in rapidly growing *Escherichia coli*. *Front Microbiol* 6:636.

32. Burmann BM, *et al*. (2010) A NusE:NusG complex links transcription and translation. *Science* 328:501–504.

33. Proshkin S, Rahmouni AR, Mironov A, Nudler E (2010) Cooperation between translating ribosomes and RNA polymerase in transcription elongation. *Science* 328:504–508.

34. Burmann BM, Rösch P (2011) The role of *E. coli* Nus-factors in transcription regulation and transcription:translation coupling: From structure to mechanism. *Transcription* 2:130–134.

35. Zhang Y, *et al*. (2014) DksA guards elongating RNA polymerase against ribosome-stalling-induced arrest. *Mol Cell* 53:766–778.

36. Dutta D, Shatalin K, Epshtein V, Gottesman ME, Nudler E (2011) Linking RNA polymerase backtracking to genome instability in *E. coli*. *Cell* 146:533–543.

37. Tehranchi AK, *et al*. (2010) The transcription factor DksA prevents conflicts between DNA replication and transcription machinery. *Cell* 141:595–605.

38. Fan H, *et al*. (2017) Transcription-translation coupling: Direct interactions of RNA polymerase with ribosomes and ribosomal subunits. *Nucleic Acids Res* 45:11043–11055.

39. Demo G, *et al*. (2017) Structure of RNA polymerase bound to ribosomal 30S subunit. *Elife* 6:e28560.

40. Saxena S, *et al*. (2018) *Escherichia coli* transcription factor NusG binds to 70S ribosomes. *Mol Microbiol* 108:495–504.

41. Ruff EF, *et al*. (2015) *E. coli* RNA polymerase determinants of open complex lifetime and structure. *J Mol Biol* 427:2435–2450.

42. Henderson KL, *et al*. (2017) Mechanism of transcription initiation and promoter escape by *E. coli* RNA polymerase. *Proc Natl Acad Sci USA* 114:E3032–E3040.

43. Roy S, Garges S, Adhya S (1998) Activation and repression of transcription by differential contact: Two sides of a coin. *J Biol Chem* 273:14059–14062.

44. Samanta S, Martin CT (2013) Insights into the mechanism of initial transcription in *Escherichia coli* RNA polymerase. *J Biol Chem* 288:31993–32003.

45. Winkelman JT, Chandrangsu P, Ross W, Gourse RL (2016) Open complex scrunching before nucleotide addition accounts for the unusual transcription start site of *E. coli* ribosomal RNA promoters. *Proc Natl Acad Sci USA* 113:E1787–E1795.

46. Mishanina T V, Palo MZ, Nayak D, Mooney RA, Landick R (2017) Trigger loop of RNA polymerase is a positional, not acid-base, catalyst for both transcription and proofreading. *Proc Natl Acad Sci USA* 114:E5103–E5112.

47. Bar-Nahum G, *et al*. (2005) A ratchet mechanism of transcription elongation and its control. *Cell* 120:183–193.

48. Nudler E, Mustaev A, Lukhtanov E, Goldfarb A (1997) The RNA-DNA hybrid maintains the register of transcription by preventing backtracking of RNA polymerase. *Cell* 89:33–41.

49. Kent T, Kashkina E, Anikin M, Temiakov D (2009) Maintenance of RNA-DNA hybrid length in bacterial RNA polymerases. *J Biol Chem* 284:13497–13504.

50. Toulokhonov I, Zhang J, Palangat M, Landick R (2007) A central role of the RNA polymerase trigger loop in active-site rearrangement during transcriptional pausing. *Mol Cell* 27:406–419.

51. Sydow JF, Cramer P (2009) RNA polymerase fidelity and transcriptional proofreading. *Curr Opin Struct Biol* 19:732–739.

52. Ray-Soni A, Bellecourt MJ, Landick R (2016) Mechanisms of bacterial transcription termination: All good things must end. *Annu Rev Biochem* 85:319–347.

53. Peters JM, Vangeloff AD, Landick R (2011) Bacterial transcription terminators: The RNA 3′-end chronicles. *J Mol Biol* 412:793–813.

54. Larson MH, Greenleaf WJ, Landick R, Block SM (2008) Applied force reveals mechanistic and energetic details of transcription termination. *Cell* 132:971–982.

55. Peters JM, *et al*. (2012) Rho and NusG suppress pervasive antisense transcription in *Escherichia coli*. *Genes Dev* 26:2621–2633.

56. Peters JM, *et al*. (2009) Rho directs widespread termination of intragenic and stable RNA transcription. *Proc Natl Acad Sci USA* 106:15406–15411.

57. Kriner MA, Sevostyanova A, Groisman EA (2016) Learning from the leaders: Gene regulation by the transcription termination factor Rho. *Trends Biochem Sci* 41:690–699.

58. Torres M, Balada JM, Zellars M, Squires C, Squires CL (2004) *In vivo* effect of NusB and NusG on rRNA transcription antitermination. *J Bacteriol* 186:1304–1310.

59. Santangelo TJ, Artsimovitch I (2011) Termination and antitermination: RNA polymerase runs a stop sign. *Nat Rev Microbiol* 9:319–329.

60. Vogel U, Jensen KF (1997) NusA is required for ribosomal antitermination and for modulation of the transcription elongation rate of both antiterminated RNA and mRNA. *J Biol Chem* 272:12265–12271.

61. Zellars M, Squires CL (1999) Antiterminator-dependent modulation of transcription elongation rates by NusB and NusG. *Mol Microbiol* 32:1296–1304.

62. Nudler E (2012) RNA polymerase backtracking in gene regulation and genome instability. *Cell* 149:1438–1445.

63. Zenkin N, Yuzenkova Y, Severinov K (2006) Transcript-assisted transcriptional proofreading. *Science* 313:518–520.

64. Sosunova E, Sosunov V, Epshtein V, Nikiforov V, Mustaev A (2013) Control of transcriptional fidelity by active center tuning as derived from RNA polymerase endonuclease reaction. *J Biol Chem* 288:6688–6703.

65. Borukhov S, Polyakov A, Nikiforov V, Goldfarb A (1992) GreA protein: A transcription elongation factor from *Escherichia coli*. *Proc Natl Acad Sci USA* 89:8899–8902.

66. Borukhov S, Lee J, Laptenko O (2005) Bacterial transcription elongation factors:

New insights into molecular mechanism of action. *Mol Microbiol* 55:1315–1324.

67. Feng G, Lee DN, Wang D, Chan CL, Landick R (1994) GreA-induced transcript cleavage in transcription complexes containing *Escherichia coli* RNA polymerase is controlled by multiple factors, including nascent transcript location and structure. *J Biol Chem* 269:22282–22294.

68. Hsu LM, Vo NV, Chamberlin MJ (1995) *Escherichia coli* transcript cleavage factors GreA and GreB stimulate promoter escape and gene expression *in vivo* and *in vitro*. *Proc Natl Acad Sci USA* 92:11588–11592.

69. Toulmé F, *et al*. (2000) GreA and GreB proteins revive backtracked RNA polymerase *in vivo* by promoting transcript trimming. *EMBO J* 19:6853–6859.

70. Perederina A, *et al*. (2004) Regulation through the secondary channel—Structural framework for ppGpp-DksA synergism during transcription. *Cell* 118:297–309.

71. Lemke JJ, *et al*. (2011) Direct regulation of *Escherichia coli* ribosomal protein promoters by the transcription factors ppGpp and DksA. *Proc Natl Acad Sci USA* 108:5712–5717.

72. Paul BJ, Berkmen MB, Gourse RL (2005) DksA potentiates direct activation of amino acid promoters by ppGpp. *Proc Natl Acad Sci USA* 102:7823–7828.

73. Doniselli N, *et al*. (2015) New insights into the regulatory mechanisms of ppGpp and DksA on *Escherichia coli* RNA polymerase-promoter complex. *Nucleic Acids Res* 43:5249–5262.

74. Rutherford ST, Villers CL, Lee JH, Ross W, Gourse RL (2009) Allosteric control of *Escherichia coli* rRNA promoter complexes by DksA. *Genes Dev* 23:236–248.

75. Roghanian M, Zenkin N, Yuzenkova Y (2015) Bacterial global regulators DksA/ppGpp increase fidelity of transcription. *Nucleic Acids Res* 43:1529–1536.

76. Satory D, *et al*. (2015) DksA involvement in transcription fidelity buffers stochastic epigenetic change. *Nucleic Acids Res* 43:10190–10199.

77. Gordon AJE, Satory D, Halliday JA, Herman C (2015) Lost in transcription: Transient errors in information transfer. *Curr Opin Microbiol* 24:80–87.

78. Kennell D, Riezman H (1977) Transcription and translation initiation frequencies of the *Escherichia coli lac* operon. *J Mol Biol* 114:1–21.

79. Golding I, Paulsson J, Zawilski SM, Cox EC (2005) Real-time kinetics of gene activity in individual bacteria. *Cell* 123:1025–1036.

80. Ling J, Reynolds N, Ibba M (2009) Aminoacyl-tRNA synthesis and translational quality control. *Annu Rev Microbiol* 63:71–78.

81. Zaher HS, Green R (2009) Fidelity at the molecular level: Lessons from protein synthesis. *Cell* 136:746–762.

82. Kramer EB, Farabaugh PJ (2007) The frequency of translational misreading errors in *E. coli* is largely determined by tRNA competition. *RNA* 13:87–96.

83. Vermulst M, *et al*. (2015) Transcription errors induce proteotoxic stress and shorten cellular lifespan. *Nat Commun* 6:8065.

84. Fan Y, *et al*. (2015) Protein mistranslation protects bacteria against oxidative stress. *Nucleic Acids Res* 43:1740–1748.

85. Gordon AJE, Satory D, Halliday JA, Herman C (2013) Heritable change caused by transient transcription errors. *PLoS Genet* 9:e1003595.

86. Tippin B, Kobayashi S, Bertram JG, Goodman MF (2004) To slip or skip, visualizing frameshift mutation dynamics for error-prone DNA polymerases. *J Biol Chem* 279:45360–45368.

87. Loman NJ, *et al*. (2012) Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol* 30:434–439.

88. Meacham F, *et al*. (2011) Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics* 12:451.

89. Gout JF, Thomas WK, Smith Z, Okamoto K, Lynch M (2013) Large-scale detection of *in vivo* transcription errors. *Proc Natl Acad Sci USA* 110:18584–18589.

90. James K, Gamba P, Cockell SJ, Zenkin N (2016) Misincorporation by RNA polymerase is a major source of transcription pausing *in vivo*. *Nucleic Acids Res* 45:1105–1113.

91. O'Farrell PH (1978) Suppression of defective translation and its role in the stringent response by ppGpp. *Cell* 14:545–557.

92. Bouadloun F, Donner D, Kurland CG (1983) Codon-specific missense errors *in vivo*. *EMBO J* 2:1351–1356.

93. Fredriksson Å, *et al*. (2007) Decline in ribosomal fidelity contributes to the accumulation and stabilization of the master stress response regulator σ^S upon carbon starvation. *Genes Dev* 21:862–874.

94. Drake JW (1991) A constant rate of spontaneous mutation in DNA-based microbes. *Proc Natl Acad Sci USA* 88:7160–7164.

95. Drake JW, Charlesworth B, Charlesworth D, Crow JF (1998) Rates of spontaneous mutation. *Genetics* 148:1667–1686.

96. Lee H, Popodi E, Tang H, Foster PL (2012) Rate and molecular spectrum of spontaneous mutations in the bacterium *Escherichia coli* as determined by whole-genome sequencing. *Proc Natl Acad Sci USA* 109:E2774–E2783.

97. Wielgoss S, *et al*. (2011) Mutation rate inferred from synonymous substitutions in a long-term evolution experiment with *Escherichia coli*. *G3 (Bethesda)* 1:183–186.

98. Arezi B, Hogrefe HH (2007) *Escherichia coli* DNA polymerase III ε subunit increases Moloney murine leukemia virus reverse transcriptase fidelity and accuracy of RT-PCR procedures. *Anal Biochem* 360:84–91.

99. Skasko M, *et al*. (2005) Mechanistic differences in RNA-dependent DNA polymerization and fidelity between murine leukemia virus and HIV-1 reverse transcriptases. *J Biol Chem* 280:12190–12200.

100. Baranauskas A, *et al*. (2012) Generation and characterization of new highly thermostable and processive M-MuLV reverse transcriptase variants. *Protein Eng Des Sel* 25:657–668.

101. Barak Z, Gallant J, Lindsley D, Kwieciszewki B, Heidel D (1996) Enhanced ribosome frameshifting in stationary phase cells. *J Mol Biol* 263:140–148.

102. Fu C, Parker J (1994) A ribosomal frameshifting error during translation of the *argI* mRNA of *Escherichia coli*. *Mol Gen Genet* 243:434–441.

103. Wenthzel AM, Stancek M, Isaksson LA (1998) Growth phase dependent stop codon readthrough and shift of translation reading frame in *Escherichia coli*. *FEBS Lett* 421:237–242.

104. Cheng ZF, Deutscher MP (2003) Quality control of ribosomal RNA mediated by polynucleotide phosphorylase and RNase R. *Proc Natl Acad Sci USA* 100:6388–6393.

105. Moran NA, McLaughlin HJ, Sorek R (2009) The dynamics and time scale of ongoing genomic erosion in symbiotic bacteria. *Science* 323:379–382.

106. Erie DA, Hajiseyedjavadi O, Young MC, von Hippel PH (1993) Multiple RNA polymerase conformations and GreA: Control of the fidelity of transcription. *Science* 262:867–873.

107. Moran NA, Bennett GM (2014) The tiniest tiny genomes. *Annu Rev Microbiol* 68:195–215.

108. Shigenobu S, Watanabe H, Hattori M, Sakaki Y, Ishikawa H (2000) Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. *Nature* 407:81–86.

109. Nakabachi A, *et al*. (2006) The 160-kilobase genome of the bacterial endosymbiont *Carsonella*. *Science* 314:267.

110. Lind PA, Andersson DI (2008) Whole-genome mutational biases in bacteria. *Proc Natl Acad Sci USA* 105:17878–17883.

111. McCutcheon JP, Moran NA (2011) Extreme genome reduction in symbiotic bacteria. *Nat Rev Microbiol* 10:13–26.

112. Vogel U, Jensen KF (1994) The RNA chain elongation rate in *Escherichia coli* depends on the growth rate. *J Bacteriol* 176:2807–2813.

113. Loewe L, Textor V, Scherer S (2003) High deleterious genomic mutation rate in stationary phase of *Escherichia coli*. *Science* 302:1558–1560.

114. Bull HJ, Lombardo M, Rosenberg SM (2001) Stationary-phase mutation in the bacterial chromosome: Recombination protein and DNA polymerase IV dependence. *Proc Natl Acad Sci USA* 98:8334–8341.

115. Bhamre S, Gadea BB, Koyama CA, White SJ, Fowler RG (2001) An aerobic *recA*-, *umuC*-dependent pathway of spontaneous base-pair substitution mutagenesis in *Escherichia coli*. *Mutat Res* 473:229–247.

116. Bridges BA (1993) Spontaneous mutation in stationary-phase *Escherichia coli* WP2 carrying various DNA repair alleles. *Mutat Res* 302:173–176.

117. Bridges BA (1996) Mutation in resting cells: The role of endogenous DNA damage. *Cancer Surv* 28:155–167.

118. Foster PL (2005) Stress responses and genetic variation in bacteria. *Mutat Res* 569:3–11.

119. Nair S, Finkel SE (2004) Dps protects cells against multiple stresses during stationary phase. *J Bacteriol* 186:4192–4198.

120. Schellhorn HE, Hassan HM (1988) Transcriptional regulation of *katE* in *Escherichia coli* K-12. *J Bacteriol* 170:4286–4292.

121. Lombardo MJ, Aponyi I, Rosenberg SM (2004) General stress response regulator RpoS in adaptive mutation and amplification in *Escherichia coli*. *Genetics* 166:669–680.

122. Leong P, Hsia HC, Miller JH (1986) Analysis of spontaneous base substitutions generated in mismatch-repair-deficient strains of *Escherichia coli*. *J Bacteriol* 168:412–416.

123. Cupples CG, Miller JH (1989) A set of *lacZ* mutations in *Escherichia coli* that allow rapid detection of each of the six base substitutions. *Proc Natl Acad Sci USA* 86:5345–5349.

124. Coulondre C, Miller JH (1977) Genetic studies of the *lac* repressor. IV. Mutagenic specificity in the *lacI* gene of *Escherichia coli*. *J Mol Biol* 117:577–606.

125. Frederico LA, Kunkel TA, Shaw BR (1990) A sensitive genetic assay for the detection of cytosine deamination: Determination of rate constants and the activation energy. *Biochemistry* 29:2532–2537.

126. McNulty DE, Claffee BA, Huddleston MJ, Kane JF (2003) Mistranslational errors associated with the rare arginine codon CGG in *Escherichia coli*. *Protein Expr Purif* 27:365–374.

127. Buckstein MH, He J, Rubin H (2008) Characterization of nucleotide pools as a function of physiological state in *Escherichia coli*. *J Bacteriol* 190:718–726.

128. Abbondanzieri EA, Greenleaf WJ, Shaevitz JW, Landick R, Block SM (2005) Direct observation of base-pair stepping by RNA polymerase. *Nature* 438:460–465.

129. Larson MH, Landick R, Block SM (2011) Single-molecule studies of RNA polymerase: One singular sensation, every little step it takes. *Mol Cell* 41:249–262.

130. Jiang Z, *et al*. (2013) Comparative analysis of genome sequences from four strains of the *Buchnera aphidicola* Mp endosymbion of the green peach aphid, *Myzus persicae*. *BMC Genomics* 14:917.

131. Stead MB, *et al*. (2012) RNAsnap^{TM}: A rapid, quantitative and inexpensive, method for isolating total RNA from bacteria. *Nucleic Acids Res* 40:e156.

132. Keseler IM, *et al*. (2013) EcoCyc: Fusing model organism databases with systems biology. *Nucleic Acids Res* 41:D605–D612.

133. Yuzenkova Y, *et al*. (2014) Control of transcription elongation by GreA determines rate of gene expression in *Streptococcus pneumoniae*. *Nucleic Acids*

Res 42:10987–10999.

134. Trautinger BW, Jaktaji RP, Rusakova E, Lloyd RG (2005) RNA polymerase modulators and DNA repair activities resolve conflicts between DNA replication and transcription. *Mol Cell* 19:247–258.

135. Washburn RS, Gottesman ME (2011) Transcription termination maintains chromosome integrity. *Proc Natl Acad Sci USA* 108:792–797.

136. Rosenberger RF, Hilton J (1983) The frequency of transcriptional and translational errors at nonsense codons in the *lacZ* gene of *Escherichia coli*. *Mol Gen Genet* 191:207–212.

137. Wagner LA, Weiss RB, Driscoll R, Dunn DS, Gesteland RF (1990) Transcriptional slippage occurs during elongation at runs of adenine or thymine in *Escherichia coli*. *Nucleic Acids Res* 18:3529–3535.

138. Parks AR, *et al*. (2014) Bacteriophage λ N protein inhibits transcription slippage by *Escherichia coli* RNA polymerase. *Nucleic Acids Res* 42:5823–5829.

139. Coenye T, Vandamme P (2005) Characterization of mononucleotide repeats in sequenced prokaryotic genomes. *DNA Res* 12:221–233.

140. Gur-Arie R, *et al*. (2000) Simple sequence repeats in *Escherichia coli*: Abundance, distribution, composition, and polymorphism. *Genome Res* 10:62–71.

141. Goldberg AL (2003) Protein degradation and protection against misfolded or damaged proteins. *Nature* 426:895–899.

142. Calloni G, *et al*. (2012) DnaK functions as a central hub in the *E. coli* chaperone network. *Cell Rep* 1:251–264.

143. Tamas I, *et al*. (2008) Endosymbiont gene functions impaired and rescued by polymerase infidelity at poly(A) tracts. *Proc Natl Acad Sci USA* 105:14934–14939.

144. Quan S, Zhang N, French S, Squires CL (2005) Transcriptional polarity in rRNA operons of *Escherichia coli nusA* and *nusB* mutant strains. *J Bacteriol* 187:1632–1638.

145. Epshtein V, Toulmé F, Rahmouni AR, Borukhov S, Nudler E (2003) Transcription through the roadblocks: The role of RNA polymerase cooperation. *EMBO J* 22:4719–4727.

146. Larson MH, *et al*. (2014) A pause sequence enriched at translation start sites drives

transcription dynamics *in vivo*. *Science* 344:1042–1047.

147. Sidorenkov I, Komissarova N, Kashlev M (1998) Crucial role of the RNA:DNA hybrid in the processivity of transcription. *Mol Cell* 2:55–64.

148. Baptiste BA, Jacob KD, Eckert KA (2015) Genetic evidence that both dNTP-stabilized and strand slippage mechanisms may dictate DNA polymerase errors within mononucleotide microsatellites. *DNA Repair* 29:91–100.

149. Kobayashi S, Valentine MR, Pham P, O'Donnell M, Goodman MF (2002) Fidelity of *Escherichia coli* DNA polymerase IV. Preferential generation of small deletion mutations by dNTP-stabilized misalignment. *J Biol Chem* 277:34198–34207.

150. Pomerantz RT, Temiakov D, Anikin M, Vassylyev DG, McAllister WT (2006) A mechanism of nucleotide misincorporation during transcription due to template-strand misalignment. *Mol Cell* 24:245–255.

151. Kashkina E, *et al*. (2006) Template misalignment in multisubunit RNA polymerases and transcription fidelity. *Mol Cell* 24:257–266.

152. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9:357–359.

153. Schirmer M, D'Amore R, Ijaz UZ, Hall N, Quince C (2016) Illumina error profiles: Resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics* 17:125.

154. Komissarova N, Kashlev M (1997) Transcriptional arrest: *Escherichia coli* RNA polymerase translocates backward, leaving the 3' end of the RNA intact and extruded. *Proc Natl Acad Sci USA* 94:1755–1760.

155. Orlova M, Newlands J, Das A, Goldfarb A, Borukhov S (1995) Intrinsic transcript cleavage activity of RNA polymerase. *Proc Natl Acad Sci USA* 92:4596–4600.

156. Yuzenkova Y, Zenkin N (2010) Central role of the RNA polymerase trigger loop in intrinsic RNA hydrolysis. *Proc Natl Acad Sci USA* 107:10878–10883.

157. Borukhov S, Sagitov V, Goldfarb A (1993) Transcript cleavage factors from *E. coli*. *Cell* 72:459–466.

158. Borukhov S, Laptenko O, Lee J (2001) *Escherichia coli* transcript cleavage factors GreA and GreB: Functions and mechanisms of action. *Methods Enzymol* 342:64–76.

159. Laptenko O, Lee J, Lomakin I, Borukhov S (2003) Transcript cleavage factors

GreA and GreB act as transient catalytic components of RNA polymerase. *EMBO J* 22:6322–6334.

160. Traverse CC, Ochman H (2017) Genome-wide spectra of transcription insertions and deletions reveal that slippage depends on RNA:DNA hybrid complementarity. *MBio* 8:e01230-17.

161. Gout JF, *et al*. (2017) The landscape of transcription errors in eukaryotic cells. *Sci Adv* 3:e1701484.

162. Erie DA, Hajiseyedjavadi O, Young MC, von Hippel PH (1993) Multiple RNA polymerase conformations and GreA: Control of the fidelity of transcription. *Science* 262:867–783.

163. Fish RN, Kane CM (2002) Promoting elongation with transcript cleavage stimulatory factors. *Biochim Biophys Acta* 1577:287–307.

164. Churchman LS, Weissman JS (2011) Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature* 469:368–373.

165. Larson MH, *et al*. (2014) A pause sequence enriched at translation start sites drives transcription dynamics *in vivo*. *Science* 344:1042–1047.

166. Imashimizu M, *et al*. (2015) Visualizing translocation dynamics and nascent transcript errors in paused RNA polymerases *in vivo*. *Genome Biol* 16:98.

167. Bubunenko MG, *et al*. (2017) A *Cre* transcription fidelity reporter identifies GreA as a major RNA proofreading factor in *Escherichia coli*. *Genetics* 206:179–187.

168. Sosunov V, *et al*. (2003) Unified two-metal mechanism of RNA synthesis and degradation by RNA polymerase. *EMBO J* 22:2234–2244.

169. Paul BJ, *et al*. (2004) DksA: A critical component of the transcription initiation machinery that potentiates the regulation of rRNA promoters by ppGpp and the initiating NTP. *Cell* 118:311–322.

170. Potrykus K, *et al*. (2006) Antagonistic regulation of *Escherichia coli* ribosomal RNA *rrnB* P1 promoter activity by GreA and DksA. *J Biol Chem* 281:15238–15248.

171. Vinella D, Potrykus K, Murphy H, Cashel M (2012) Effects on growth by changes of the balance between GreA, GreB, and DksA suggest mutual competition and functional redundancy in *Escherichia coli*. *J Bacteriol* 194:261–273.

172. Gamba P, James K, Zenkin N (2017) A link between transcription fidelity and

pausing *in vivo*. *Transcription* 8:99–105.

173. Miller JH (1972) Experiments in Molecular Genetics. *Cold Spring Harbor Laboratory Press*, Cold Spring Harbor, NY.